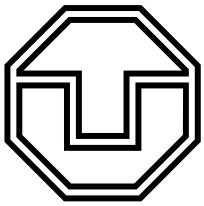

TECHNISCHE UNIVERSITÄT DRESDEN

Universality and variability in the statistics of data with fat-tailed distributions: the case of word frequencies in natural languages

Dissertation
zur Erlangung des wissenschaftlichen Grades
Doctor rerum naturalium

vorgelegt von

Martin Gerlach
geboren am 17.07.1984 in Dresden

erstellt am
Max-Planck-Institut für Physik komplexer Systeme
Dresden
2015

Eingereicht am 22. Oktober 2015
Verteidigt am 01. März 2016

Gutachter:
Prof. Dr. Jan-Michael Rost
Prof. Dr. Roland Ketzmerick
Prof. Alvaro Corral

Abstract

Natural language is a remarkable example of a complex dynamical system which combines variation and universal structure emerging from the interaction of millions of individuals. Understanding statistical properties of texts is not only crucial in applications of information retrieval and natural language processing, e.g. search engines, but also allow deeper insights into the organization of knowledge in the form of written text. In this thesis, we investigate the statistical and dynamical processes underlying the co-existence of universality and variability in word statistics. We combine a careful statistical analysis of large empirical databases on language usage with analytical and numerical studies of stochastic models. We find that the fat-tailed distribution of word frequencies is best described by a generalized Zipf's law characterized by two scaling regimes, in which the values of the parameters are extremely robust with respect to time as well as the type and the size of the database under consideration depending only on the particular language. We provide an interpretation of the two regimes in terms of a distinction of words into a finite core vocabulary and a (virtually) infinite noncore vocabulary. Proposing a simple generative process of language usage, we can establish the connection to the problem of the vocabulary growth, i.e. how the number of different words scale with the database size, from which we obtain a unified perspective on different universal scaling laws simultaneously appearing in the statistics of natural language. On the one hand, our stochastic model accurately predicts the expected number of different items as measured in empirical data spanning hundreds of years and 9 orders of magnitude in size showing that the supposed vocabulary growth over time is mainly driven by database size and not by a change in vocabulary richness. On the other hand, analysis of the variation around the expected size of the vocabulary shows anomalous fluctuation scaling, i.e. the vocabulary is a nonself-averaging quantity, and therefore, fluctuations are much larger than expected. We derive how this results from topical variations in a collection of texts coming from different authors, disciplines, or times manifest in the form of correlations of frequencies of different words due to their semantic relation. We explore the consequences of topical variation in applications to language change and topic models emphasizing the difficulties (and presenting possible solutions) due to the fact that the statistics of word frequencies are characterized by a fat-tailed distribution. First, we propose an information-theoretic measure based on the Shannon-Gibbs entropy and suitable generalizations quantifying the similarity between different texts which allows us to determine how fast the vocabulary of a language changes over time. Second, we combine topic models from machine learning with concepts from community detection in complex networks in order to infer large-scale (mesoscopic) structures in a collection of texts. Finally, we study language change of individual words on historical time scales, i.e. how a linguistic innovation spreads through a community of speakers, providing a framework to quantitatively combine microscopic models of language change with empirical data that is only available on a macroscopic level (i.e. averaged over the population of speakers).

Zusammenfassung

Natürliche Sprache ist ein bemerkenswertes Beispiel eines komplexen dynamischen Systems, welches Variation und universelle Struktur, die aus der Interaktion von Millionen von Individuen hervorgehen, kombiniert. Ein Verständnis der statistischen Eigentschaften von Texten ist nicht nur wichtig in Anwendungen der Informationsrückgewinnung (*information retrieval*) und Sprachverarbeitung (*natural language processing*), z.B. in Suchmaschinen, sondern erlaubt ebenfalls tiefere Erkenntnises über die Organisation von Wissen in Form von Texten. In dieser Thesis untersuchen wir die statistischen und dynamischen Prozesse, welche der Koexistenz von Universalität und Variabilität in der Statistik von Wörtern zu Grunde liegen. Wir kombinieren eine sorgfältige statistische Analyse von grossen Mengen empirischer Daten über Sprachgebrauch mit analytischen und numerischen Studien stochastischer Modelle. Wir stellen fest, dass die *fat-tail* Verteilung von Worthäufigkeiten am besten durch ein generalisiertes Zipf'sches Gesetz, charakterisiert durch zwei Skalierungsregime, beschrieben wird. Dabei sind die Werte der Parameter extrem robust bezüglich der Zeit, sowie des Typs und der Größe der untersuchten Daten und hängen lediglich von der jeweiligen Sprache ab. Wir präsentieren eine Interpretation der zwei Skalierungsregime in Form eines endlichen Kernwortschatzes und eines (praktisch) unendlich großen Nicht-Kernwortschatz. Mit Hilfe eines simplen generativen Prozesses können wir die Verbindung zum Problem des Wachstums der Wortschatzgröße herstellen, d.h. wie die Anzahl verschiedener Wörter mit der Gesamtanzahl an Wörtern (der Menge an Daten) skaliert, wodurch wir eine vereinheitlichte Perspektive auf verschiedene universelle Skalierungsgesetze, welche gleichzeitig in der Statistik natürlicher Sprache auftreten, erhalten. Einerseits erlaubt unser stochastisches Modell akkurate Vorhersagen über die Anzahl verschiedener Elemente in empirisch gemessenen Daten, welche aus einer Zeitspanne von mehreren hundert Jahren stammen und sich über 9 Größenordnungen spannen. Dies verdeutlicht, dass das vermeintliche Wachstum des Wortschatzes mit der Zeit hauptsächlich durch eine Zunahme der Datenmenge bedingt ist und nicht durch eine Zunahme des Reichtums des Wortschatzes. Andererseits zeigt die Analyse der Variation um die erwartete Wortschatzgröße sogenannte anomale Skalierung der Fluktuationen, d.h. die Wortschatzgröße ist nicht selbst-mittelnd, sodass Fluktuationen viel größer sind als erwartet. Wir zeigen wie diese Beobachtungen aus topischen Variationen in einer Kollektion von Texten aus verschiedenen Zeiten, Disziplinen oder von verschiedenen Autoren resultieren, welche sich in Form von Korrelationen von Häufigkeiten zwischen verschiedenen Wörtern (auf Grund ihrer semantischen Beziehung) manifestieren. Wir untersuchen die Konsequenzen dieser topischen Variationen in Anwendungen bezogen auf Sprachwandel sowie *topic models*. Dabei gehen wir insbesondere auf die Schwierigkeiten (und mögliche Lösungen) ein in Anbetracht der Tatsache, dass die Statistik von Worthäufigkeiten durch eine *fat-tail* Verteilung charakterisiert ist. Erstens schlagen wir ein informationstheoretisches Maß vor, welches auf der Shannon-Gibbs Entropie und geeigneten Generalisierungen beruht, um die Ähnlichkeit zwischen verschiedenen Texten zu quantifizieren. Dies ermöglicht uns zu ermitteln, wie schnell sich der Wortschatz einer Sprache mit der Zeit verändert. Zweitens kombinieren wir *topic models* aus dem Bereich maschinellem Lernens mit dem Konzept *community detection* in komplexen Netzwerken für die Inferenz mesoskopischer Struktur in einer Kollektion von Texten. Schließlich untersuchen wir Sprachwandel an Hand einzelner Wortfor-

men auf einer historischen Zeitskala, d.h. wie sich eine linguistische Innovation in einer Population von Sprechern verbreitet. Wir formulieren einen Ansatz in dem wir quantitativ mikroskopische Modelle von Sprachwandel mit empirischen Daten, welche typischerweise nur auf einer makroskopischen Ebene verfügbar sind (d.h. als Mittelwert über eine Population), kombinieren.

Titel: Universalität und Variabilität in der Statistik von Daten mit *fat-tail* Verteilungen: der Fall von Worthäufigkeiten in natürlichen Sprachen.

Contents

1. Introduction

In this thesis, we study natural languages as a remarkable example of complex dynamical systems which combines variation and universal structure emerging from the interaction of millions of individuals. Following the paradigm of complex systems, i.e. that complex patterns can often be understood as the result of simple rules, we use tools from statistical physics and non-linear dynamics to show how simple models are able to capture main statistical features and help to understand the underlying dynamical processes. In Sec. 1.1, we sketch the main ideas involved in this endeavor, in particular how collective human behaviour can be studied in the framework of complex systems, how this builds on fundamental ideas developed in physics, especially statistical physics, and how this can be applied to studying natural language. In Sec. 1.2 we define the scope of this thesis and give a brief description of the individual chapters.

1.1. Complex systems, physics, and language

From a historical perspective, the (quantitative) modeling of collective social behaviour by physicists can be traced back to the very beginnings of modern statistical physics as formulated by Maxwell and Boltzmann. In fact, it was not merely the application of ideas from physics to explain social phenomena, but rather a mutual interaction, e.g. as described in a historical account given in Ref. [Bal04]: *"Today physicists regard the application of statistical mechanics to social phenomena as a new and risky venture. Few, it seems, recall how the process originated the other way around, in the days when physical science and social science were the twin siblings of a mechanistic philosophy and when it was not in the least disreputable to invoke the habits of people to explain the habits of inanimate particles."* These early quantitative approaches in the analysis on the statistical regularities in empirical data resulting from human behaviour are exemplified by the seminal works of i) Pareto on the distribution of income [Par96]; ii) Auerbach, a theoretical physicist, on the distribution of city sizes [Aue13]; iii) Lotka measuring scientific productivity [Lot26]; or iv) Estoup and Zipf on the statistics of words in natural language [Est16, Zip36]. However, it was not before the second half of the 20th century until the emergence of a mathematical formalization of how these macroscopic observations (e.g. statistical regularities) could be explained and derived from microscopic theories (the interaction of individuals). This was achieved by combining concepts from critical phenomena in the description of phase transitions, self-organization (e.g. the paradigmatic model of self-organized criticality [BTW87]) in the description of non-equilibrium phenomena, as well as chaos in dynamical systems. The underlying paradigm can be best summarized by the phrase "more is different" [And72], expressing the notion that many macroscopic phenomena can *only* be understood by considering the mutual interactions

among their constituent parts (in contrast to a reductionist view focusing on the properties of the constituent itself), that is they are emergent properties. This view was further corroborated by the idea of complex networks [AB02, New10], which studies the structure of the local interactions in terms of their (often non-trivial) topology. This leads to models that are often simplistic descriptions of social systems, however, the idea of universality [Sta99] implies that many large-scale phenomena do not depend on the microscopic details of the underlying process, thus offering a unifying perspective on seemingly unrelated phenomena including, e.g., the behaviour and flocking of crowds, spreading of information or diseases among the population, or the distribution of attention paid to different items (i.e. economy of attention) [CFL09]. Following this line of thought, empirical observations of scaling laws were interpreted as signatures of emergent universal behaviour and were a main driving factor in the study of complex systems [New05, Sor06], especially with the formulation of the idea of scale-free networks [BA99]. In this context, the role of empirical data, and that of statistics more generally, becomes of crucial importance. While the recent availability of large electronic records on human activity (even on the microscopic scale), e.g. on the Internet, were a main driving factor in the formulation and development of theoretical models, they have to be directly confronted with a careful statistical analysis of empirical data. This is necessary in order to not only corroborate perpetual claims of universality, but also to assess the empirical support of the proposed models and their predictive power [SP12] in the spirit of statistician G. Box [Box79]: *'Essentially, all models are wrong but some are useful"*. However, as of today *"there is a striking imbalance between empirical evidence and theoretical modeling, in favor of the latter"* [CFL09].

In this perspective, one can study natural language as an example of a complex system [BBB$^+$09], i.e. it can be considered as an emergent property from the interaction of millions of individuals. More generally, it reflects human activities and interests and serves as a marker of external events and how humans influence each other. Furthermore, it offers unique opportunities in terms of abundant availability of data in the form of written text: not only due to the increase of contemporary usage of language, e.g. on the Internet, but also on historical time scales due to preservation and digitization of books from the past, e.g. the Google-ngram database [MSA$^+$11]. Such an approach provides an understanding on the statistical properties of texts and how human interests and language itself changes over time. In addition, insights on the underlying structure of language are also crucial in applications of statistical natural language processing [MS99], e.g. in the field of information retrieval [MRS08], in particular search engines [CMS09], but also more generally in the problem of how human knowledge is organized in the form of written texts [SB04].

In the view of complex systems, one tries to uncover and understand the universal structure observed in the statistics of word usage. These regularities are often expressed in the form of universal scaling laws. One of the fundamental concepts in the analysis of natural languages builds on the Shannon entropy formulated in the framework of information theory. In strong analogy to the Gibbs-Boltzmann entropy from statistical physics, the Shannon entropy quantifies the degree of randomness contained in an instance of text.

Most notably, the fat-tailed distribution of word frequencies, the so-called Zipf's law, has been

shown to exist in virtually any instance of written text, yet its origin is still highly debated [Pia14]. In his seminal work, Zipf himself claimed that this regularity is a result of the *principle of least effort* [Zip49], i.e. that natural languages evolve in a way that they are optimized such that speakers and hearers can communicate efficiently. The idea of a trade-off between speakers and hearers was formalized in Ref. [FS03] in an information-theoretic model which exhibits a phase transition, where only at the critical point the system allows for meaningful communication in which the word frequency distribution resembles that of natural languages. These findings sparked much interest in the form of extensions and generalizations, but also initiated an ongoing discussion on the viability of the model and the interpretation of its results [PAOP10, CMFS11, DMA12]. Other approaches investigating statistical regularities in the form of universal scaling laws concern, e.g., the hierarchical structure in the organization of language. Starting from the works of Mandelbrot [DM73], written texts considered as a time series were shown to exhibit long-range correlations from highly structured linguistic levels down to the building blocks of texts (words and even individual characters) [Gra89, EP94, ALDEM06, ACE12]. These intricate signatures are far from being understood and are crucial in understanding how humans use language to mirror their activities and experiences from the natural world.

Despite the fact that language exhibits a remarkable degree of universal structure, it is constantly subject to transformation leading to language change over time [WLH68, Cro00]. In the more general context of cultural evolution, it constitutes an example of how social conventions emerge in a group of individuals [CB15]. Recent efforts aim at understanding these dynamical processes from microscopic models of interaction in the framework of complex systems [BBB$^+$09]. One notable approach in this line of thought is the so-called utterance selection model [BBCM06], a mathematical model of language change in the form of a many-body stochastic process taking into account evolutionary aspects and the non-trivial topology of social networks. While this allows insights into qualitative features of language change, e.g. based on general arguments of existing symmetries in the model [BC12], a quantitative comparison to empirical data is difficult [BBCM09] since data on historical time scales is scarce and typically not available on the level of individual speakers.

A more exhaustive overview on the different branches within the intersection of complex systems, physics, and natural language can be obtained from the bibliography [Bib] compiled by the author of this thesis.

1.2. Scope of this thesis

In this thesis, we investigate the statistical and dynamical processes underlying the co-existence of universality and variability in word statistics. Motivated by the recent availability of data on language usage unprecedented in size we re-examine previously reported empirical "universal laws", e.g. Zipf's and Heaps' law. Combining data analysis with analytical and numerical investigations of stochastic models we provide a unifying perspective on the appearance of different universal laws. While on average, these laws are extremely robust across different databases and even languages, on closer inspection, fluctuations around these laws are much larger than expected due to the variability of word

frequencies across texts from different authors, topics, or times. We show how these variations can be systematically exploited for i) quantifying the similarity between different instances of text such that one can measure, e.g. how fast the vocabulary of a language changes over time; or ii) inferring large-scale structures in a collection of texts allowing for the identification of groups (clusters) of semantically related documents. In this, we emphasize the difficulties (and present possible solutions) encountered in applications to (necessarily) finite data resulting from the fact that the statistics of word frequencies is characterized by a fat-tailed distribution. We further study language change of individual words, i.e. linguistic innovations, on historical time scales providing a framework to quantitatively combine microscopic models of language change with empirical data that is only available on a macroscopic level (i.e. averaged over the population of speakers).

We start in Ch. 2 by introducing the basic concepts (i.e. scaling, fat-tailed distributions, and more generally the statistical analysis of natural language) employed in this thesis. In Ch. 3 we assess the extent of universality in the fat-tailed distribution of word frequencies finding two scaling regimes whose parameters are remarkably robust across different times as well as databases and only depend on the language. In the framework of sampling processes we interpret this result in terms of a core vocabulary and explore the implications of this scaling on the problem of the vocabulary growth. In Ch. 4 we investigate the variability in the distribution of word frequencies, i.e. the unequal dissemination of words across different documents due to topical variation. On the one hand, we show that this leads to much larger fluctuations than are typically expected from simple null models in the example of the vocabulary growth. On the other hand, we illustrate how topical variations can be used to quantify the (dis-) similarity of texts from different authors, disciplines, or times and discuss the problems of finite-size estimations encountered in the presence of the fat-tailed distribution of word frequencies. In Ch. 5 we approach the problem of modeling the variability in word frequencies in terms of identifying coherent groups or topics (i.e. large-scale structures) in written text. In this, we employ methods developed in the field of community detection in complex networks and show how this yields a more general formulation of the problem and leads to better results, as well as an improved understanding, compared to current state-of-the-art methods. In Ch. 6 we consider the variability of word frequencies in the description of language change over time. Taking advantage of our previous results, we quantify how fast the vocabulary of a language is changing on historical time scales. On the level of individual words, we investigate how new words considered as linguistic innovations spread through a community of speakers. A summary and a discussion of the main results is presented in Ch. 7. Appendix A describes in detail the different databases of written language and how they were obtained and filtered. In Appendix B we provide additional evidence for the generality of the results obtained in Sec. 3.1 by repeating the respective analysis for different languages and databases. Appendix C provides analytical and numerical arguments for how to rescale the vocabulary growth in the presence of a threshold discussed in Sec. 3.2. In Appendix D we discuss a possible extension of the Jensen-Shannon divergence introduced in Sec. 4.3.

This thesis is a result of studies performed at the Max Planck Institute for the Physics of Complex Systems between 2012 and 2015 under the supervision of Dr. Eduardo G. Altmann. The scientific

results presented in this thesis are contained in Refs. [GA13, GA14, GGMA14, GFCA15, GPA15, AG16], which are cited in the introduction of the corresponding chapter.

2. Basic concepts

2.1. Scaling laws and fat-tailed distributions

2.1.1. Scaling laws in complex systems

In this section we briefly sketch the idea of scaling laws and their paradigmatic character in the context of complex systems.

A complex system is typically seen as composed of interacting parts that display emergent behaviour, iconically captured by the notion that "more is different" [And72]. In lack of a more precise (agreed upon) definition the appearance of scaling laws in general, and power laws in particular, is often seen as a characteristic feature of complex systems [New11b, Mit11]. We say that an observable $y(x)$ is scale-invariant (i.e. it "scales") under the transformation $x \mapsto ax$, if there exists $b(a)$ such that [Sor06]

$$y(ax) = b(a)y(x). \tag{2.1}$$

This defines a homogeneous function and is solved by a power law,

$$y(x) = Cx^{\alpha} \tag{2.2}$$

with $\alpha = \ln b / \ln a$ (this concept can be generalized to universal scaling functions if y depends on more than one variable, see e.g. [SDM01]). Therefore, if in a system we find a scaling law of the form Eq. (2.2), the observable y is said to lack a characteristic scale in x.

Many different natural and social systems have been reported to show such power-law scalings, see [Mit04, New05, CSN09] for recent reviews. In fact, the analysis of such systems in terms of scaling laws can be dated back to the end of the 19-th century to the works of i) Pareto on the uneven distribution of income [Par96]; ii) Auerbach, a German physicist, on the power-law distribution of city sizes [Aue13]; iii) Estoup [Est16] and Zipf [Zip49] on the distribution of word frequencies; iv) Arrhenius on the relation between the number of species found in a habitat of a given area [Arr21]; or v) Kleiber, who investigated a non-trivial scaling between body size and the metabolic rate of animals [Kle47], also known as allometric scaling [Wes97]. Note that, while the latter example is a relation between two variables, the other examples refer to the distribution of a single variable, thus, distinguishing two types of scaling laws.

A big motivation for studying such diverse systems in the form scaling laws came from the notion of universality in analogy to the theory of critical phenomena in statistical physics [Sta99, SDM01,

Sor06]. The main idea is that for systems near the critical point (the point of, e.g., a 2nd-order phase transition), macroscopic phenomena do not show a characteristic scale (e.g. diverging correlation lengths). Furthermore, macroscopically these systems show a strikingly similar behaviour despite the fact that they are otherwise quite different in nature. Using tools from statistical physics (e.g. renormalization group) it can be formally shown that in the critical regime the properties of large-scale phenomena are independent of the microscopic details of the underlying processes offering a unifying perspective on seemingly unrelated phenomena (i.e. universality classes) [SDM01]. However, in the framework of critical phenomena, this holds if the respective control parameter is "tuned" to its critical value, limiting its direct applicability to, e.g., social phenomena. One solution to this obstacle was put forward in terms of the notion of self-organized criticality (SOC) [BTW87]. There, it was shown how simple non-equilibrium systems could naturally end up in a critical point by means of self-organization (e.g. the sandpile model). Although SOC can be considered one of the most stimulating concepts in offering a unified description of the observed scaling phenomena in different natural, social, or biological systems, even 25 years after its inception, there is still extensive debate on the applicability to real systems beyond its paradigmatic character [WPC+15]. In addition, it has been stressed that power laws are not only generated by SOC, but originate from a multitude of processes [Mit04, New05, Sor06]. Therefore, it is self-evident that empirical evidence of scaling laws cannot be taken as a proof of SOC as the underlying dynamical process. Furthermore, the empirical evidence of the supposedly ubiquitous appearance of power laws in natural and social systems has been questioned recently in terms of their validity [CSN09, SP12] or whether other functions provide a better description of the data, e.g. for the case of the city-size distribution [Eec04, Lev09, Eec09]. While the statistical methods employed in assessing the validity of power laws are not free of choices and can be questioned themselves, see e.g. [DC13], even if the exact functional form is not given by a power law, the widespread appearance of non-trivial scaling relations has important practical consequences. Besides their peculiar statistical properties when dealing with finite data (e.g. divergent values for the mean or standard deviation [BG90]) these concern, e.g., i) the prediction of extreme events in the context of, for example, floods, financial crisis, epileptic seizures [AJK06], or attention [MA14]; or ii) the sub- or super-linear scaling of socio-economical indicators (such as GDP, patents, CO_2 emissions) with city size [BLH+07, Bat13] potentially useful in urban planning [LB14].

2.1.2. Fat-tailed distributions

In this section we apply the notion of scaling laws to probability distributions of a random variable.

We consider a continuous, non-negative random variable $X \in [0, \infty)$ with probability mass function (pmf) $p(x)$, i.e. the probability that $X \in [x, x + \mathrm{d}x)$ is given by $P(x \leq X < x + \mathrm{d}x) = p(x)\mathrm{d}x$, with the normalization condition $\int_x \mathrm{d}x p(x) = 1$. Introducing the m-th moment as

$$\langle x^m \rangle \equiv \int_0^\infty \mathrm{d}x x^m p(x), \tag{2.3}$$

a pmf is said to be fat-tailed if $\exists \alpha > 0$ such that

$$\begin{aligned} &\forall m < \alpha : \langle x^m \rangle < \infty, \\ &\forall m \geq \alpha : \langle x^m \rangle = \infty, \end{aligned} \tag{2.4}$$

i.e. the m-th moment is only defined for $m < \alpha$. A general class of pmf's that satisfy this condition is given by power-law distributions of the form

$$p(x) \sim x^{-(1+\alpha)}, \tag{2.5}$$

where $A(x) \sim B(x)$ indicates that $\lim_{x\to\infty} A(x)/B(x) = \text{constant}$. Thus they decay as a power law with exponent $\alpha+1$ for $x \gg 1$ showing the scaling behaviour discussed in Sec. 2.1.1 (at least) asymptotically.

We note that fat-tailed distributions are a subset of so-called heavy-tailed distributions, that is distributions that decay slower than exponential, which is a less strict statement than Eq. (2.4). One example of a distribution that decays slower than an exponential (heavy-tailed) but faster than a power law (not fat-tailed) is the log-normal distribution. We further mention that fat-tailed distributions appear naturally in the context of the generalized central limit theorem [Nol15] and extreme value theory [AJK06]. The generalized central limit theorem due to Gnedenko and Kolmogorov states that the distribution of a sum of independent and identically distributed random variables converges to a so-called stable distribution. If the variance of the individual random variables is infinite (finite) the resulting stable distribution shows the scaling behaviour of Eq. (2.5) with $\alpha \in (0, 2)$ (is a Gaussian distribution). In extreme value theory, the Fisher–Tippett–Gnedenko theorem states that the distribution of maxima of a sequence of independent and identically distributed random variables converges to one of three limiting distributions: the Gumbel-, Fréchet-, or Weibull-distribution. The Fréchet-distribution shows the scaling behaviour of Eq. (2.5) with $\alpha > 0$.

Finally, we want to illustrate the connection to so-called rank-frequency distributions. Considering a countable set of items denoted by $i = 1, \ldots, S$ (with S possibly infinite) each with probability of occurrence p_i such that $\sum_{i=1}^{S} p_i = 1$, we construct the rank-frequency distribution $f(r)$ by assigning a rank $r = 1, \ldots, S$ to each item i, i.e. $r = r_i$ and defining $f(r) = f(r_i) = p_i$ such that $f(r) \geq f(r')$ for $r < r'$. In other words, the probabilities of the items i are ordered in decreasing order and the items are mapped to positive integers (the ranks). In this view, the random variables drawn from the distribution $f(r)$ are the ranks r. In the limit of large ranks, i.e. $r \gg 1$, we approximate the frequencies f themselves as a continuous random variable X with pmf $p(f)$ defined as above, Eq. (2.4), i.e. $P(f \leq X < f + \mathrm{d}f) = p(f)\mathrm{d}f$. Noting that the rank r_i simply counts the number of items that have a frequency larger (or equal) than $f(r_i) = p_i$ we can write

$$r(f) \propto P(X \geq f) \tag{2.6}$$

which corresponds to the complementary cumulative distribution defined by

$$P(X \geq f) \equiv \int_f^1 \mathrm{d}f' p(f') \tag{2.7}$$

such that we can relate the rank-frequency distribution to $p(f)$ by

$$p(f) \propto \frac{\mathrm{d}r(f)}{\mathrm{d}f}. \tag{2.8}$$

If we assume a power law rank-frequency distribution for the ranks $r = 1, 2, \ldots, \infty$

$$f(r) \sim r^{-\gamma} \tag{2.9}$$

with $\gamma > 1$ we get that

$$p(f) \sim f^{-(1+1/\gamma)} \tag{2.10}$$

establishing the relation $\alpha = 1/\gamma$ with respect to fat-tailed distributions considered in Eq. (2.5). Note that the latter scaling for $p(f)$ only holds for $0 < f \ll 1$ due to i) the singularity at $f = 0$; and ii) the scaling in $f(r)$ for $r \gg 1$ implies $f \ll 1$.

2.2. Statistical analysis of language

In this section we briefly introduce the basic terminology in the statistical analysis of natural language.

The object of study are language corpora, i.e. collections of written texts produced by speakers of a language. The texts consist of a sequence of characters defined by the corresponding alphabet. Here, we analyze the texts on the level of words[1] (or 1-grams), which we take as any string of characters delimited by white spaces (a string of n successive words is then called an n-gram). At this point, it is useful to distinguish the term word into what is called a word-*type* and a word-*token*. A word-type identifies any unique string of characters, i.e. the set of word-types corresponds to the set of distinct words in a text. In contrast, a word-token refers to each individual occurrence of a given word-type in a text.

The statistical analysis of language finds wide application in problems from computer science (in this context it is called statistical natural language processing [MS99]) involving, e.g. (automatic) classification of documents or authorship recognition. In the following we illustrate two different approaches in analyzing the structure of language in a statistical framework.

Name of the law	Observables	Functional form
Zipf [Zip36, Pia14]	f: frequency of word w; r: rank of w in f	$f(r) = \alpha r^{-\beta}$
Menzerath-Altmann [Alt80, Cra05]	x: length of the whole; y: size of the parts	$y = \alpha x^{\beta} e^{-\gamma x}$
Heaps [Her64, Hea78, Egg07]	V: number of words; N : database size	$V \propto N^{\alpha}$
Recurrence [Zip49, APM09, LPPM11]	τ: distance between words	$P(\tau) \propto \exp{(\alpha\tau)^{\beta}}$
Long-range correlation [DM73, SZZ93, ACE12]	$C(\tau)$: autocorrelation at lag τ	$C(\tau) \propto \tau^{-\alpha}$
Entropy Scaling [EP94, Deb06]	H: Entropy of text with blocks of size n	$H \propto \alpha n^{\beta} + \gamma n$
Information content [Zip36, Zip49, PTG11]	$I(l)$: Information of word with length l	$I(l) = \alpha + \beta l$
Taylor's law [GA14]	σ: standard deviation around the mean μ	$\sigma \propto \mu^{\alpha}$
Networks [SCMVS09, CM09, BFPS+13, CL14]	Topology of lexical/semantic networks	various

Table 2.1.: List of linguistic laws. Examples of linguistic laws with a non-exhaustive list of references to the literature, a brief description of the observables involved in the linguistic law, and the functional form stated by the law.

2.2.1. Linguistic laws

One aim in the statistical analysis of words (and text constituents more generally) is to reveal regularities in the usage of language in the form of linguistic laws (following the notation from quantitative linguistics [KAP05, Baa01, Zan14]), see Tab. 2.1 for a (non-exhaustive) list of examples.

While some of the linguistic laws clearly intend to speak about the language as whole, in practice they are tested and motivated by observations in specific texts which are thus implicitly or explicitly assumed to reflect the language as a whole. In particular, these laws denote quantitative relationships between measurements obtained in a written text or corpus in contrast to, e.g. syntactic rules. The distinction to rules is crucial insofar as the existence of these laws do not directly affect the production of (grammatically and semantically) meaningful sentences in a language (e.g., because they involve scales of the text much larger or shorter than a sentence). It is thus not difficult to get convinced that a creative and persistent daemon[2], trained in the techniques of constrained writing [Wik14a], can generate understandable and arbitrary long texts which deliberately violate any single law mentioned above. In a strict Popperian sense, a single of such demonic texts would be sufficient to falsify the proposed linguistic law. Linguistic laws are thus different from syntactic rules and require a different interpretation, in which statistics (and probability theory) play a central role, in contrast to the laws of classical Physics. In the framework of a statistical interpretation, the demonic texts mentioned above would be considered untypical, or highly unlikely. However, this requires the formulation of a generative (stochastic) process for the production of texts in order to quantify the expected fluctuations. In the same sense that a scientific law cannot be judged separated from a theory, linguistic laws are only fully defined once a generative process is given. The crucial role of generative processes in the statistical analysis of language was already emphasized by Gustav Herdan, the founding father of quantitative linguistics [Her64]: *"The quantities which we call statistical laws being only expectations, they are subject to random fluctuations whose extent must be regarded as part of the statistical law."*

[1] In Ref. [Zan14] it is argued that "the overall organization of language is more related to the distribution and ordering of words than to the arrangement of letters ... [because] the coding of words in a particular alphabet or phoneme set is, to a large extent, irrelevant to the linguistic structure of communication."

[2] a relative of Maxwell's daemon known from thermodynamics.

2.2.2. Information theory

Another approach in the statistical analysis of language aims at quantifying the amount of information contained in language thereby measuring its degree of order. This question was a major stirring force already in Shannon's seminal paper [Sha48] on developing a mathematical theory of communication which can be considered the birth of the field of information theory. The main idea is a probabilistic description in which one assumes the existence of a stochastic source emitting symbols $i = 1, 2, \ldots, S$ randomly with probabilities $\boldsymbol{p} = (p_1, p_2, \ldots, p_S)$ with $\sum_{i=1}^{S} p_i = 1$. This defines a stochastic process in which the symbols are random variables. The central quantity in the framework of information theory is the entropy of the source defined as

$$H(\boldsymbol{p}) \equiv -\sum_{i=1}^{S} p_i \log p_i, \tag{2.11}$$

where the unit of H is bits if one identifies $\log = \log_2$ as the logarithm to base 2. It quantifies the average amount of uncertainty about a randomly drawn symbol from the source, that is the (average) number of binary questions one has to ask in order to learn the identity of the symbol. For example, the case of maximum uncertainty is encountered if all symbols have the same probability, i.e. $p_i = 1/S$, which gives $H_{\max} = \log S$, whereas in the case where we are absolutely sure about which symbol will be drawn, i.e. there exists one symbol i^* for which $p_{i^*} = 1$ and $p_{i \neq i^*} = 0$, we get $H_{\min} = 0$ (note that $\lim_{x \to 0^+} x \log x = 0$). For a thorough treatment of information-theoretic measures based on the entropy, Eq. (2.11), we refer to Ref. [CT06].

In applications to language, one considers texts as sequences of symbols, e.g. letters or words, which are realizations of a stochastic source and assigns each symbol a probability by counting the number of occurrences in the observed sequences. In this framework, however, the obtained entropy only quantifies the uncertainty about a source which we assume to emit symbols randomly without taking into account any correlations induced by the ordering of the symbols present in the original sequence. The latter can be achieved by extending the definition of a symbol in the sequence considering not only individual letters or words (i.e. 1-grams) but also strings of n letters or words (i.e. n-grams) as individual symbols and taking the limit $n \to \infty$. In practice, one considers $n \gg 1$ due to the finitude of available data.

Beyond applications in language, we note that the concepts of information theory can be applied much more generally to quantify and analyze the regularities in any set of data, e.g. in the framework of the Minimum Description Length [Gr7].

2.3. Linguistic databases

In this section we briefly describe the databases we employ in the statistical analysis of natural language. For details on, e.g. how the data was obtained or how it was filtered, we refer to Appendix A.

The *Google-ngram* database [MSA+11] constitutes one of the largest electronic and publicly avail-

able collections of written text comprising millions of books published in the period 1520-2008. It provides the timeseries of word frequencies with a yearly resolution, i.e. the number of times a word occurs in the books published in a given year. Our main interest in this database stems from its large size (several millions of books with $> 10^{11}$ words) and from the long time span it covers thus enabling us to trace historical changes in the usage of language. For each language we use two different partitions of the database: i) yearly (y), in which case $y(t)$ corresponds to the database of the year t; and ii) cumulative (Y), in which case $Y(t) = \sum_{t'=to}^{t} y(t')$. Despite its large size, the Google-ngram has some severe limitations, e.g. i) the data is already coarse-grained as the frequency of a word in a given year constitutes an average taken over many books; ii) the collection of data is affected by errors in the (automatic) scanning and digitization of books, so-called OCR-errors; or iii) it is unknown which books are included in the dataset of a given year, possibly leading to biased samples [PDD15a].

In order to address these issues and to emphasize that any findings obtained from the Google-ngram are of general validity, we choose two additional databases which still contain enough data but do not suffer from these drawbacks, i.e. i) where the statistics of words can be analyzed on the level of individual texts; ii) the text does not need to be scanned; and iii) the publishing process is inherently different from that of books. The *Wikipedia* database contains all articles of a complete snapshot of the English Wikipedia [Wik] ($3,743,306$ articles and $\gtrsim 10^9$ word-tokens). The *PlosOne* database contains all articles published in the journal PlosOne [API] ($76,723$ articles and $\approx 10^9$ word-tokens).

3. Scaling laws as a sampling problem

Statistics of word usage share remarkable similarities with other social, physical, and biological systems. The most well-known similarity is the widespread appearance of fat-tailed distributions, e.g. Zipf's law which shows that words in a text span a wide range of frequencies. These regularities are often expressed in the form of *scaling laws* and seem to be extremely robust with respect to language, topic, and time, despite the fact that all languages are constantly changing and individual word frequencies show a strong variation across time and topics. The recent availability of large databases allows us to reach new levels of quantitative precision and opens up possibilities of making new analyses. This requires the application of rigorous statistical methods, also in order to test the validity of the purported scaling laws in empirical data.

In this chapter we analyze scaling laws in the framework of stochastic *sampling* processes as simple models for the usage of words and, hence, the production of texts. These models reveal the basic mechanisms that govern, e.g. the vocabulary growth, and show the connection between the different scaling laws observed simultaneously. Since our models are formulated in a probabilistic and generative manner they can be viewed as simple null models allowing for the assessment of the validity of the respective scaling laws. They have the additional significance as a remarkable example of how simple models are able to capture the main statistical features that emerge from the interaction of millions of individuals (or components).

In the following sections we combine statistical analysis of written texts and simple stochastic models to explain the appearance of and elucidate the connection between scaling laws in the statistics of word frequencies [GA13, AG16]. In Sec. 3.1 we perform a careful analysis of the rank-frequency distribution of words proposing a generalization to Zipf's law. In Sec. 3.2 we show how this affects the vocabulary growth, i.e. how many different words will be found as a function of the total text length, also known as Heaps' law.

3.1. Distribution of word frequencies: Zipf's law

In this section we focus on the distribution of word frequencies. In his seminal work, Zipf proposed that if we rank all word-types according to the frequency of appearance ($r = 1, 2, \ldots, V$), the frequency $f(r)$ of the r-th word-type scales with the rank r as [Zip36, Zip49]

$$f(r) = \frac{f(1)}{r}, \tag{3.1}$$

where $f(1)$ is the frequency of the most frequent word. The above expression cannot hold for large r because for any $f(1) > 0$, there is an r^* such that $\sum_{r=1}^{r^*} f(1)/r > 1$ meaning that $f(r)$ has to decay faster than $1/r$ for $r \gtrsim r^*$.[1] Taking also into account that $f(1)$ may not be the best proportionality factor, the modern version of Zipf's law is

$$f(r) = Cr^{-\gamma}, \tag{3.2}$$

with $\gamma \geq 1$, motivating numerous different generalization of Zipf's proposal [Man53, Tul96, BBM11]. While many of these proposals were shown to provide a better account of particular databases, they remain in a great extent unsatisfactory because they lack the simplicity and universality of Zipf's original proposal (e.g., the parameters vary depending on the size, topic or date of publication of the analyzed texts [CMH97, Fer05]).

Motivated by the unprecedented magnitude of recently available databases, we apply rigorous statistical tests to determine which of the previously proposed distributions provide a better account of the data. We will give an overview on the previously proposed models in Sec. 3.1.1. In Sec. 3.1.2 we describe the details of the statistical analysis in terms of i) fitting, ii) model selection, and iii) hypothesis testing. In Sec. 3.1.3, we apply this framework to the Google-ngram database in 5 different languages (English, German, French, Spanish, and Russian), see Appendix A.1 for details of the data. In Sec. 3.1.4 we conclude with some critical remarks on the intricacies encountered in the statistical analysis of Zipf's law in particular, and fat-tailed distributions in general.

3.1.1. Models

We select 7 of the most popular previously proposed heavy-tailed distributions with at most two free parameters [Baa01, LMC10, JÏ2]:

1. Power law (P):

$$f(r;\gamma) = Cr^{-\gamma} \tag{3.3}$$

2. Shifted power law (SP):

$$f(r;\gamma,b) = C(r+b)^{-\gamma} \tag{3.4}$$

3. Power law with exponential cutoff, tail (PET):

$$f(r;\gamma,b) = C\exp(-br)\, r^{-\gamma} \tag{3.5}$$

4. Power law with exponential cutoff, beginning (PEB):

$$f(r;\gamma,b) = C\exp(-b/r)\, r^{-\gamma} \tag{3.6}$$

[1] In English $f(1) \approx 0.07$ (the frequency of "the") such that $\sum_{r=1}^{r^*} f(1)/r > 1$ for $r^* \approx 10^6$.

5. Log-normal (LN):

$$f(r;\mu,\sigma) = Cr^{-1}\exp\left(-\frac{1}{2}\left(\ln r-\mu\right)^2/\sigma^2\right) \tag{3.7}$$

6. Weibull (W):

$$f(r;\gamma,b) = Cr^{\gamma-1}\exp\left(-br^{-\gamma}\right) \tag{3.8}$$

7. Double power law (DP):

$$f(r;\gamma,b) = C\begin{cases} r^{-1}, & r\leq b \\ b^{\gamma-1}r^{-\gamma} & r>b, \end{cases} \tag{3.9}$$

The notation $f(r;\theta)$ means that the distribution f depends on the rank r, and θ is the set of parameters. The normalization constant $C = C(\theta)$ is a function of the respective parameters and fixed by $\sum_{r=1}^{\infty} f(r;\theta) = 1$. In practice, this is calculated numerically with the Euler-Maclaurin formula available in the package mpmath [J+10].

In the following, we answer the following three questions:

- What are the best parameters for each model? (*fitting*)
- Which model is more likely to describe the data? (*model selection*)
- What is the probability that the data was generated by each model? (*validity*)

3.1.2. Statistical methods

Fitting

Given the data (i.e. the number of occurrences of each word), the best set of parameters, θ^*, is that which minimizes the likelihood of the model [HTF09]. The need for such Maximum Likelihood (ML) methods when fitting power-law distributions (such as Zipf's law) has been emphasized in many recent publications [GMY04, Bau07]. We refer to the review article Ref. [CSN09] and references therein for more details on the (by now well-established) methods for fitting fat-tailed distributions.

In practice, it is convenient to minimize the negative of the log-likelihood:

$$\theta^* = \arg\min_{\theta}\mathcal{L}'\left(\theta\right), \tag{3.10}$$

where the likelihood $\mathcal{L}$ is simply the probability of the data given the model with parameters θ such that we can write

$$\mathcal{L}'\left(\theta\right) = -\ln\mathcal{L}\left(\theta\right) = -\sum_{i=1}^{N}\ln f\left(r\left(i\right);\theta\right). \tag{3.11}$$

where we assume that each word-token i (and hence its corresponding rank $r(i)$) is drawn independently from the corresponding distribution $f(r;\theta)$. In practice, the minimization is performed numerically with a Nelder-Mead simplex algorithm (available in the Scipy library [JOP$^+$]).

Model selection

Given the best parameters θ^* obtained from fitting each model to the data, we determine which of the proposed models $i = 1...7$ is more likely to describe the data. This is done by comparing their likelihoods through the log-likelihood ratio [BA02]. For two models m_1 and m_2 with fitted parameters $\theta^*_{m_1}$ and $\theta^*_{m_2}$ the value $\log \mathcal{L}_{m_1}(\theta*_{m_1})/\mathcal{L}_{m_2}(\theta*_{m_2}) = 1(-1)$ means it is $e^1 = 2.718...$ times more (less) likely that the data was generated by model m_1 than model m_2. When comparing models with different numbers of parameters, one may penalize more complex models (i.e. models with a larger number of parameters) in order to avoid overfitting. Therefore, instead of comparing the likelihoods directly, we calculate the Akaike information criterion (AIC) [Aka74] for each model i

$$AIC_i = 2\mathcal{L}'_i(\theta^*) + 2R_i, \tag{3.12}$$

where R_i is the number of parameters estimated in the model i. The model which gives the minimum value $AIC_{\min} = \min_i\{AIC_i\}$ is most likely to describe the given data. From this we can calculate the relative likelihood l_i [BA02]

$$l_i = e^{(-(AIC_i - AIC_{\min})/2)}, \tag{3.13}$$

which states how likely model i is to describe the data in comparison with the best model. This implies that the probability w_i that model i (out of the 7 models considered) describes the data is given by [BA02]

$$\tilde{l}_i = l_i / \sum_{j=1}^{7} l_j. \tag{3.14}$$

Validity

In the last step, we want to test the hypothesis that the data was generated from each of the models. The idea is to quantify the similarity between the observed data and the data we would obtain if the proposed model (with the fitted parameters θ^*) was true by means of the p-value. A low p-value (e.g., p-value < 0.01) is a strong indication that the null hypothesis (the model) is violated.

In practice [CSN09], one defines a measure of distance D between the data and the model (e.g. the Kolmogorov-Smirnov distance) and estimates the p-value as the fraction of finite-size realizations of the proposed model (i.e. sets of randomly sampled data from the model with parameters θ^* with the same number of observations) that show a distance $D' > D$. However, due to the large size of our data ($N > 10^{11}$), this becomes computationally unfeasible, since it would require us to draw $\approx 10^{15}$ random numbers in order to obtain a p-value precision of 0.01.

Therefore, we calculate an approximate p-value from the corresponding χ^2-statistics [D'A86]:

$$\chi^2 = \sum_{j=1}^{Q} = \frac{(N_j - n_j)^2}{n_j}. \tag{3.15}$$

Here the domain is partitioned into Q cells, such that the expected number of observations per cell $n_j \geq 5$ [Tay97], with N_j being the actual observed number of observations in cell j.

3.1.3. Results

Here we show our detailed analysis for the largest database (Google-ngram English, $t_0 = 1520$, $t \in [1805, 2000]$). For the other 4 languages we report the main findings and leave the details for the Appendix B.

The results show that it is extremely unlikely ($p < 10^{-15}$) that the data was drawn exactly from any of the proposed distributions, a consequence of the large databases which makes any small (true) deviation incompatible with these simple fits. On the other hand, the results show unequivocally that for English the distribution with two power laws (DP) is the best fit ($1-\tilde{l}_{\rm dp} < 10^{-15}$) for all databases with a size larger than 10^9 words. In Fig. 3.1 we show the AIC of the proposed models, Eqs. (3.3-3.9). We confirm that the double power law is also the best fit for a completely independent database, the English Wikipedia, a strong indication of the validity of this result in databases of different origin (see Appendix B).

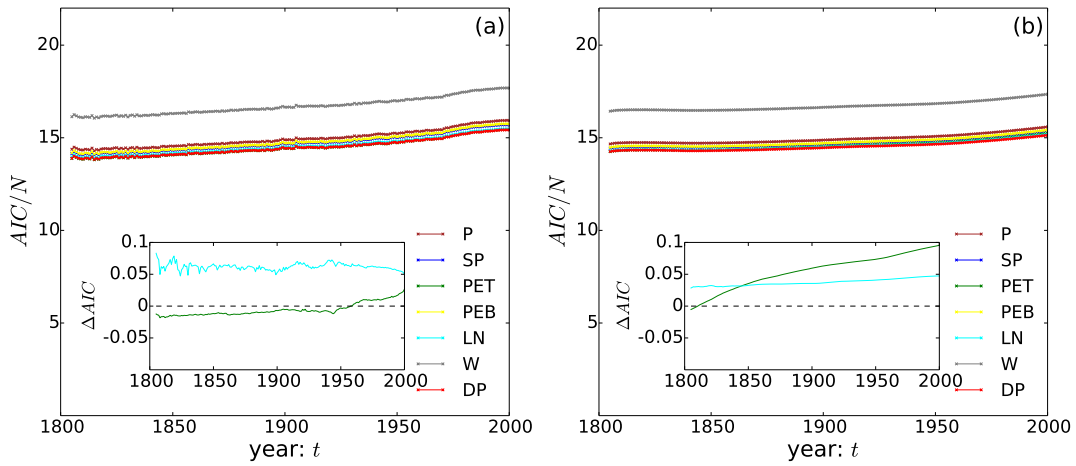

Figure 3.1.: Discrimination between different models with AIC for English. Value of the AIC for a) yearly data $y(t)$ b) cumulative data $Y(t)$. The inset shows the difference $\Delta AIC = AIC_{\rm i}/N - AIC_{\rm DP}/N$, $i \in \{\mathrm{P}, \mathrm{SP}, \mathrm{PET}, \mathrm{PEB}, \mathrm{LN}, \mathrm{W}\}$ meaning that if $\Delta AIC > 0$ the double power law is the most likely model among the proposed describing the data. Numbers refer to the enumeration of the model in Sec. 3.1.1.

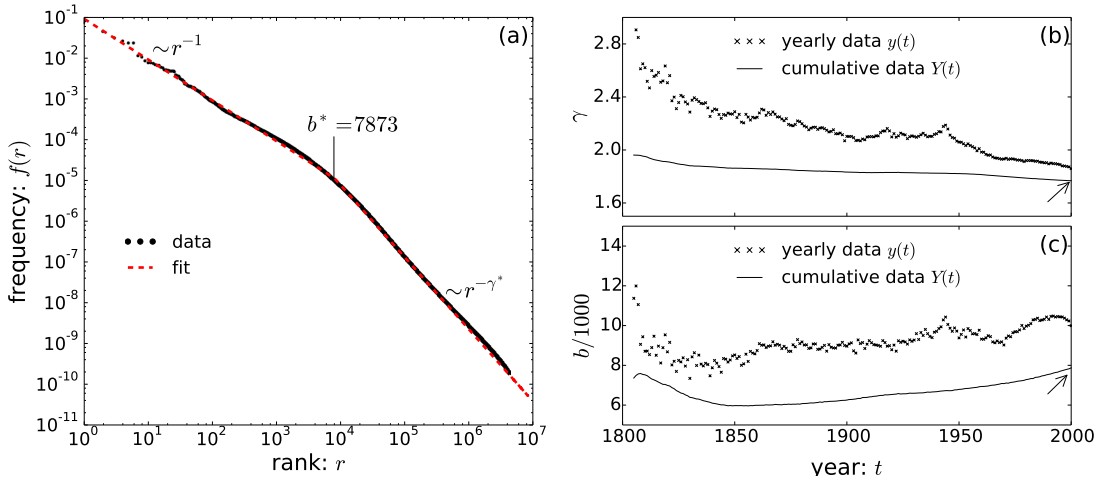

Figure 3.2.: Rank-frequency distribution shows double scaling behavior (Zipf's plot). a) Rank-frequency distribution for the English database $Y(2000)$ (solid) and a ML-fit of Eq. (3.16) (dashed). b+c) parameters γ and b obtained from ML-fits of Eq. (3.16) to yearly $y(t)$ (x-symbols) and accumulated $Y(t)$ (solid) database. Arrows indicate the values of the parameters γ^* and b^* obtained for the fit in a). Results are shown for the time range $t \in [1805, 2000]$ in which data is most reliable, accumulation starts in $t_0 = 1520$.

We now discuss in detail the best two-parameter model we identify from our data, see Fig. 3.2(a):

$$f_{\rm dp}(r;\gamma,b) = C \begin{cases} r^{-1}, & r \le b \\ b^{\gamma-1} r^{-\gamma} & r > b, \end{cases} \tag{3.16}$$

characterizing a double power law (DP), where b, and γ are free parameters, and $C = C(\gamma, b)$ is the normalization constant[2]. The original Zipf's law, Eq. (3.1), is recovered for high-frequency words and a critical rank $r = b$ determines a transition to a power law with exponent γ. Double power laws were proposed as a generalization of Zipf's law in Ref. [NB98] and further investigated in Refs. [FS01, PTH+12]. These insightful works used distributions with two power-law exponents γ_1, γ_2 and were motivated by the visual inspection of double logarithmic plots. Our improved statistical analysis confirms and extends these observations for the simpler distribution Eq. (3.16). Besides the likelihood analysis and visual inspection given in Fig. 3.2, a third strong evidence in favor of distribution (3.16) comes from the comparison of the estimated parameters of different corpora shown in Fig. 3.2(b,c). Very similar values $b \in [7 \cdot 10^3, 12 \cdot 10^3]$ and $\gamma \in [1.8, 2.5]$ were obtained for non-overlapping databases (for the English Wikipedia: $b = 7830$, $\gamma = 1.68$), and the fluctuations become smaller for increasing database size. These observations strongly suggest that the same fixed parameters provide a good description of all English texts (e.g., $y(1900)$ and $y(2000)$). Therefore,

[2]The sharp transition between the two regimes in Eq. (3.16) might seem artificial. We believe that alternative distributions which interpolate between the two scalings could provide a similarly good account of the data. The advantage of the distribution Eq. (3.16) is that the transition point $r = b$ appears explicitly as a free parameter and can be independently estimated from data

hereafter we do not consider individual fits for each database and instead assume that Eq. (3.16) is valid with $b = b^* = 7873$ and $\gamma = \gamma^* = 1.77$, values obtained for our largest database $Y(2000)$.

Similar findings also apply to the other languages. In Tab. 3.1 we summarize the parameters γ^* and b^* obtained from a ML-fit of the largest database $Y(2000)$ of the respective language to Eq. (3.16). French and Spanish are also best described by Eq. (3.16) for databases exceeding a particular size and yield values for γ^* and b^* similar to English. For German and Russian Eq. (3.16) constitutes only the second best model. However, we have strong indications that it provides a better account of the tails ($r \gg b^*$) and therefore we expect that even larger databases will reveal the double power law as the best fit also in these languages (see Appendix B). Apart from being the smallest databases among the investigated languages the large values of b^* in German and Russian require even larger databases to characterize the deviations from the r^{-1} regime for $r \gg b^*$.

Language	b^*	γ^*	$C^* = C(\gamma^*, b^*)$
English	7,873	1.77	0.0922
French	8,208	1.78	0.0920
Spanish	8,757	1.78	0.0915
German	19,863	1.62	0.0828
Russian	62,238	1.94	0.0789

Table 3.1.: Parameters b^*, γ^*, and $C^* = C(\gamma^*, b^*)$ obtained from ML-fit of Eq. (3.16) obtained for the largest database $Y(2000)$ for all considered languages.

3.1.4. Critical discussion on fitting

In this section we want to critically discuss the use of Maximum Likelihood approaches in the statistical analysis of the distribution of word frequencies, e.g. Zipf's law. These remarks are equally valid in the analysis of statistical laws in social or natural systems in general. The statistical analysis is far from being free of choices, both in terms of the methods employed and also about additional assumptions not contained in the original law. The main aim is to point out that these choices matter and should be carefully discussed.

Representation matters

Instead of analyzing the rank-frequency distribution $f(r)$, an alternative formulation can be obtained by looking at the fraction of word-types with a given frequency f, $P(f)$. In this representation, Zipf's law, Eq. (3.2), can be cast in the form

$$P(f) = C^\dagger f^{-\gamma^\dagger}. \tag{3.17}$$

While asymptotically this formulation of Zipf's law is equivalent to the one in terms of ranks $f(r)$ with $\gamma^\dagger = 1 + \gamma$ [Man61, Mit04, New05], the likelihood computed in both cases is usually not the same. The reason is that real data is finite, noisy, and possibly not drawn from this distribution.

Power-law fitting manuals [CSN09] – employed for linguistic and non-linguistic problems – suggest to fit Zipf's law using the distribution of frequencies $P(f)$ given in Eq. (3.17). However, it is also possible to use the rank formulation, Eq. (3.2) because the frequencies of ranks $f(r)$ are normalized $\sum_r f(r) = 1$ and can thus be interpreted as a probability distribution. However, a drawback in fitting $f(r)$ is that the process of ranking introduces a bias in the estimator [GLSW96, MSFCC15]. For instance, consider a finite sample from a true Zipf distribution containing ranks $r = 1, \ldots, \infty$. Because of statistical fluctuations, some of the rankings will be inverted (or absent) so that when we rank the words according to the observations we would obtain observations (the rank) different from the ones drawn. This effect introduces bias in our estimation of the parameters (overestimating the quality of the fit). The words affected by this bias are the ones with largest ranks, which contribute very little to the estimation of the parameters of Zipf's law. Therefore, we expect that this bias to become negligible for sufficiently large sample sizes.

In the likelihood of $P(f)$ each observation corresponds to the frequency of a word-*type* meaning that the most frequent word in the database (e.g., *the*) counts the same as words appearing only once (the hapax legomena). This means that, in practice, the part of the distribution that matters in the fitting are the words with very few counts, which contribute very little to the total text. Instead, in the likelihood of $f(r)$ the observational quantity is the rank of each occurrence of the word meaning that each word-*token* counts the same. Thus, the frequent words contribute more to the likelihood, a good property that we believe makes the fitting more robust.

Therefore, the statistical rigorous methods of Maximum Likelihood will be dominated either by the most frequent (in case of fitting in $f(r)$) or least frequent (in case of fitting in $P(f)$) words. As a result, even if asymptotically (i.e. infinite data) different formulations of a statistical law are equivalent, the specific representation in which we test the law implicitly assumes a certain generative (stochastic) process how the data was sampled. This in turn leads to different results when applied to finite and often noisy data, which has to be taken into account when interpreting the results of the corresponding fits. In our opinion, the advantage of fitting the rank-frequency distribution is that it is more robust since every individual word-token in the text is counted as a separate observation, i.e. words that appear more often (and whose frequency is more stable) contribute more to the likelihood function.

Correlated samples

Typically, more data confirms the original observations motivating the statistical laws – mostly based on visual inspection – but tends to make these laws violate any rigorous statistical test designed to evaluate their validity. This is seen in Fig. 3.2(a), where a visually good fit yields a p-value $< 10^{-15}$. This leads to a seemingly contradictory situation: while the validity of the laws as an estimation of the general behavior is confirmed (e.g., it is much better than alternative descriptions), these laws are strictly speaking falsified.

The failure of passing significance tests for increasing data size is not surprising because any small deviation from the null model becomes statistically significant. The conclusion emerging from these

analysis is that power-law distributions are not as widely valid as previously claimed (see also, e.g. [CSN09, SP12]), but often are better than alternative (simple) descriptions, e.g. considering two-parameter generalizations of Zipf's law, Eq. (3.16). Our main criticism on that view is that it ignores the presence of correlations in the data: the computation of the likelihood in Eq. (3.11) assumes independent observations. Furthermore, this assumption leads to an underestimation of the expected fluctuations (e.g. KS-distance) in the calculation of the p-value when assessing the validity of the statistical law. It is thus unclear in which extent a negative result in the validity test (e.g., p-value$\ll 0.01$) is due to a failure of the proposed law or, instead, is due to the violation of the hypothesis of *independent* sampling. This hypothesis is known to be violated in texts [Baa01, LPPM11] – the sequence of words and letters are obviously related to each other. In Fig. 3.3 we show that these correlations affect the estimation of the frequency of individual words, which show fluctuations much larger than those expected when assuming independence.

One approach to take into account correlations is to estimate a time for which two observations are independent, and then consider observations only after this time (a smaller effective sample size). Alternative approaches considered statistical tests for specific classes of stochastic processes (correlated in time) [Wei78] or based on estimations of the correlation coming from the data [CB11]. The application of these methods to linguistic laws is not straightforward because these methods fail in cases in which no characteristic correlation time exists. Books show such long-range correlations [DM73, SZZ93], also in the position of individual words in books [DM73, ACE12], in agreement with the observations reported in Fig. 3.3. More generally, correlations lead to a slower convergence to asymptotic values and it is thus possible to create processes of text generation which comply to a linguistic law asymptotically but that (in finite samples) violate statistical tests based on independent

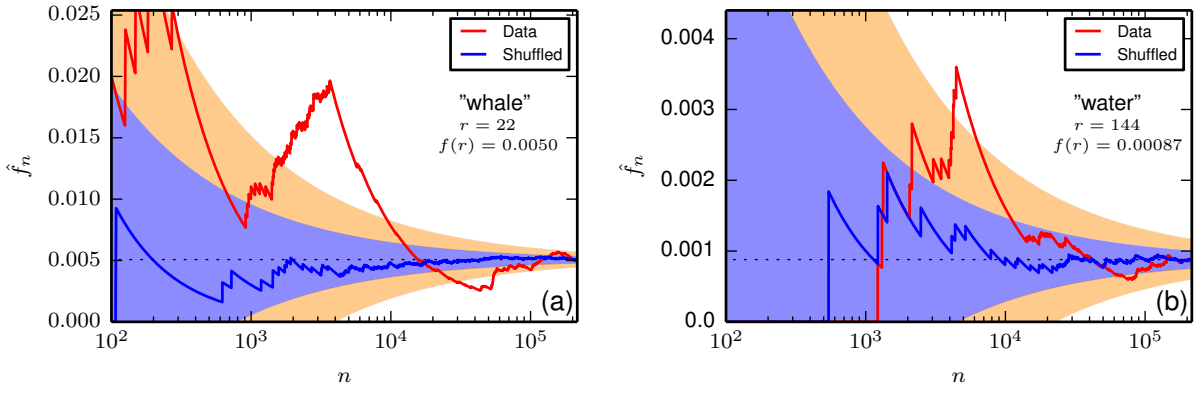

Figure 3.3.: Estimation of the frequency, $\hat{f}_n$ of a word in the first n word-tokens of a book (Moby Dick by H. Melville). The red curve corresponds to the actual observation and the blue curve to the curve measured in a version of the book in which all word-tokens were randomly shuffled. The shaded regions show the expected fluctuations ($\pm 2\sigma$) assuming that the probability of using the word is given by the frequency of the word in the whole book ($f_{n=N}$) and that: (i) usage is random (blue region) – see also Ref. [Baa01] or (ii) the time between successive usages of the word is drawn randomly from a stretched exponential distribution with exponent $\beta = 0.5$, as proposed in Ref. [APM09].

sampling. While this problem also affects the fitting and, thus, the model comparison, the effect of correlations affects all models in the same way which allows for a comparison on equal foot. We, therefore, believe that the results from the model comparison are more robust with respect to these correlations justifying our previous approach.

3.2. Vocabulary growth: Heaps' law

Another well-studied scaling in language concerns the vocabulary growth and is known as Heaps' law [Hea78] (although already proposed much earlier by Herdan [Her58]). It states that the number of different words (word-types), V, scales sub-linearly with the total number of words (word-tokens), N, i.e.

$$V(N) \propto N^{\lambda} \text{ for } N \gg 1, \tag{3.18}$$

with $0 < \lambda < 1$.

The importance of looking at this scaling law is that it is used in different applications [MRS08], e.g., (i) to optimize the memory allocation in inverted indexing algorithms [BYN00], e.g., used in search engines [WZ05, CMS09]; (ii) to estimate the vocabulary of a language [MSA+11, Kle13]; or (iii) to compare the vocabulary richness of documents with different lengths [WA99, Baa01, YKK12]. Beyond linguistic applications, scalings of the number of unique items as a function of database size similar to Heaps' law have been observed in other domains, e.g. the species-area relationship in ecology [Arr21, Bra82, GG06], collaborative tagging [CBB+09], network growth [KR13], in the statistics of chess moves [PJSB13], and in the dynamics of innovation [TLSS14].

Here, we show the intimate connection between Zipf's law and Heaps' law in the light of the results from the previous section, Sec. 3.1. For this, we propose two qualitatively different generative models in which the usage of words (and therefore the production of artificial texts) is modeled by stochastic processes. The first model in Sec. 3.2.1 explores the implications of the generalized Zipf's law on the expected vocabulary growth showing that a simple null model neglecting any correlations present in real texts allows for an accurate quantitative prediction of the vocabulary only given the rank-frequency distribution. The second model in Sec. 3.2.2 is in the same spirit of preferential attachment-type growth models offering an improved interpretation of the double scalings in our empirical findings on the generalized Zipf's law from Sec. 3.1.

3.2.1. Poisson null model

The connection between Zipf's law, Eq. (3.2), and Heaps' law, Eq. (3.18) is known at least since Mandelbrot [Man61], and has been further investigated in recent years [vv05, ZM05, SFM09], especially for large databases [WZ05], finite text sizes [BCM09, LZZ10, FCC15], and more general distributions [GA13, FCBC13]. A simple and powerful approach is the so-called Zipfian ensemble [Eli11], which can be traced [PTH+12] back to Mandelbrot [Man61]. It was shown that for a

stochastic process with fixed frequencies for words (or similar assumptions), asymptotically Heaps' law can be interpreted as a direct consequence of a Zipfian rank frequency distribution $f(r) \sim r^{-\gamma}$ [BYN00, vv05, SFM09, BCM09, Eli11] and vice versa [Sim55, ZM05, MKEHG11].

In this description, we assume that the usage of each word with rank r is governed by an independent Poisson process with a given frequency $f(r)$. As a result, the number of different words, V, becomes a stochastic variable for which we can calculate the expectation value $\mathbb{E}[V(N)]$ and the variance $\mathbb{V}[V(N)]$ over the realizations of the Poisson process:

$$\mathbb{E}[V(N)] \equiv \mu(N) = \sum_r 1 - e^{-Nf(r)}, \tag{3.19}$$

$$\mathbb{V}[V(N)] \equiv \sigma(N)^2 \equiv \mathbb{E}\left[V(N)^2\right] - \mathbb{E}[V(N)]^2 = \sum_r e^{-Nf(r)} - e^{-2Nf(r)}. \tag{3.20}$$

These expressions can be shown by noting that the number of different words in each realization of the Poisson process is given by

$$V(N) = \sum_r I\left[n_r(N, f(r))\right], \tag{3.21}$$

in which n_r is the integer number of times the word r occurs in a Poisson process of length N with frequency $f(r)$ and $I[x]$ is an indicator-type function, i.e. $I[x = 0] = 0$ and $I[x \geq 1] = 1$. Averaging over realizations of the Poisson process requires the calculation of $\mathbb{E}[I[n_r(N, f(r))]] \equiv \langle I[n_r(N)]\rangle = 1 - e^{-Nf(r)}$, which is the probability that the word with rank r appears at least once in a text of length N. Considering all words we obtain

$$\mathbb{E}[V(N)] = \sum_r \langle I[n_r(N)]\rangle = \sum_r 1 - e^{-Nf(r)}, \tag{3.22}$$

$$\mathbb{V}[V(N)] \equiv \mathbb{E}\left[V(N)^2\right] - \mathbb{E}[V(N)]^2 \tag{3.23}$$

$$= \sum_{r,r'} \langle I[n_r] I[n_{r'}]\rangle - \sum_{r,r'} \langle I[n_r]\rangle\langle I[n_{r'}]\rangle \tag{3.24}$$

$$= \sum_r \langle I[n_r]^2\rangle + \sum_{\substack{r,r' \\ r\neq r'}} \langle I[n_r] I[n_{r'}]\rangle - \sum_{r,r'} \langle I[n_r]\rangle\langle I[n_{r'}]\rangle \tag{3.25}$$

$$= \sum_r \langle I[n_r]\rangle + \sum_{\substack{r,r' \\ r\neq r'}} \langle I[n_r]\rangle\langle I[n_{r'}]\rangle - \sum_{r,r'} \langle I[n_r]\rangle\langle I[n_{r'}]\rangle \tag{3.26}$$

$$= \sum_r e^{-Nf(r)} - e^{-2Nf(r)} \tag{3.27}$$

where we used that $I[x]^2 = I[x]$ and that Poisson processes of different words ($r \neq r'$) are independent of each other.

Assuming Zipf's law (3.3), i.e. $f(r) = Cr^{-\gamma}$ with $\gamma > 1$, we can calculate $\mathbb{E}[V(N)]$ analytically in

the continuum limit by substituting $x \equiv f(r)$:

$$\begin{aligned}\mathbb{E}\left[V(N)\right] &= \sum_r 1 - e^{-Nf(r)} \\ &= -\int_0^1 \mathrm{d}x \frac{\mathrm{d}r}{\mathrm{d}x}\left(1 - e^{-Nx}\right) \\ &= \frac{1}{\gamma} C^{1/\gamma} \int_0^1 \mathrm{d}x x^{-1-1/\gamma} \mathrm{d}x \left(1 - e^{-Nx}\right) \\ &= C^{1/\gamma}\left[-1 + 1/\gamma E_{1+1/\gamma}(N) - \Gamma(-1/\gamma) N^{1/\gamma}\right]\end{aligned} \tag{3.28}$$

where $\Gamma(x)$ is the Gamma function and $E_z(x) = \int_1^\infty \mathrm{d}t e^{-xt}/t^z$ is the exponential integral with $\lim_{x\to\infty} E_z(x) = 0$ (for $z > 1$) [AS72]. Hence, for $N \gg 1$ we recover Heaps' law, Eq. (3.18), i.e. $\mathbb{E}\left[V(N)\right] \propto N^\lambda$, with a simple relation between the scaling exponents $\gamma = \lambda^{-1}$.

The Poisson null model (PNM), Eq. (3.19,3.20), therefore, is able not only to describe the connection between Zipf's law and Heaps' law in terms of the respective exponents, γ and λ, but also allows for a rigorous quantitative treatment. One the one hand, we are able to quantify the expected fluctuations around the average behaviour, which will be further explored in Sec. 4.2. On the other hand, we obtain a generalized picture in which we can predict the vocabulary growth for arbitrary rank-frequency distributions, $f(r)$.

In the following, we search for the implications of our finding of a generalized Zipf's law, Eq. (3.16), on the vocabulary growth. In order to be able to compare the predictions from the PNM to the data of Google-ngram (in which words have to appear at least $s = 41$ times before they are considered part of the vocabulary, see Appendix A.1), we have to generalize the PNM accordingly. In analogy to Eq. (3.21), we define the vocabulary as a stochastic variable

$$V^{(s)}(N) = \sum_r I_s\left[n_r(N, f(r))\right], \tag{3.29}$$

where we introduce a generalized indicator function, $I_s[x < s] = 0$ and $I_s[x \geq s] = 1$, accounting for the fact that a word with rank r has to appear at least $n_r \geq s$ times before it is included in the vocabulary. From basic properties of the Poisson process we know that

$$\mathbb{E}\left[I_s[n_r(N, f(r))]\right] \equiv \langle I_s[n_r(N)]\rangle = 1 - \sum_{j=0}^{s-1} \frac{(f(r)N)^j}{j!} e^{-f(r)N}, \tag{3.30}$$

which gives

$$\mathbb{E}\left[V^{(s)}(N)\right] = \sum_r \left[1 - \sum_{j=0}^{s-1} \frac{(f(r)N)^j}{j!} e^{-f(r)N}\right]. \tag{3.31}$$

Assuming a double power law in the rank-frequency distribution, $f_{\mathrm{dp}}(r; \gamma, b)$ from Eq. (3.16), we

can show that for $s \gg 1$ we get approximately

$$\mathbb{E}\left[V_{\mathrm{dp}}(N;\gamma,b)\right] \approx C_s \begin{cases} N, & N \ll N_b \\ N_b^{1-1/\gamma} N^{1/\gamma}, & N \gg N_b, \end{cases} \tag{3.32}$$

where N_b is the number of words such that $\mathbb{E}\left[V(N_b)\right] = b$ and the scaling constant $C_s = C/s$ [$C \approx f(r=1)$ being the frequency of the most common word, as can be seen from Eq. (3.16)]. We show this by noting that in Eq. (3.31) the term $Y \equiv 1 - \sum_{j=0}^{s-1} \frac{(f(r)N)^j}{j!} e^{-f(r)N}$, corresponds to the cumulative distribution of a Poisson distributed variable X with intensity $f(r)N$, i.e. $Y = P(X \geq s) = Q(s, f(r)N)$, where $Q(s,x) = \frac{\gamma(s,x)}{\Gamma(s)}$ with $\gamma(s,x)$ the lower incomplete Gamma function, and $\Gamma(s)$ the Gamma function. Looking at the rescaled variable $N' = N/s$, we get $P(X \geq s) = Q(s, f(r)N's)$ such that $\lim_{s\to\infty} Q(s, f(r)N's) = \Theta(1 - f(r)N')$ (see Eq. (8.11.13) in Ref. [DLM]), where $\Theta(x<0) = 0$ and $\Theta(x>0) = 1$ is the Heaviside function. Plugging this into Eq. (3.31) we get

$$\lim_{s\to\infty} \mathbb{E}\left[V^{(s)}(N' = N/s)\right] = \sum_r \Theta(N' - 1/f(r)), \tag{3.33}$$

which defines an implicit function for the vocabulary growth in the limit $s \to \infty$

$$\mathbb{E}\left[V^{(s)}(N' = 1/f(r))\right] = r \tag{3.34}$$

yielding Eq. (3.32) for the rank-frequency distribution from Eq. (3.16). In practice, Eq. (3.34) is already a good approximation for $s > 10$, see Appendix C.

In Fig. 3.4 we show that the data in the Google-ngram database obeys the scalings of Eq. (3.32). In Fig. 3.4(a) we present the $V(N)$-curve for English. While for the yearly database $y(t)$ we obtain a set of points for each t, the cumulative database $Y(t)$ builds a curve of vocabulary growth for increasing t. Despite the differences in these databases, all the data lie in a relatively narrow region of the plot which resembles a single curve compatible with the double scaling of Eq. (3.32). This curve is well described by the $\mathbb{E}[V(N)]$ curve obtained from the combination of the double power-law distribution Eq. (3.16) with fixed parameters (γ^*, b^*) and the assumption of Poisson usage of words, in the spirit of the PNM with threshold $s = 41$, Eq. (3.31). Similar observations apply to all considered languages, as shown in Fig. 3.4(b). On closer inspection, Fig. 3.4(c), the fine details of the $V(N)$ curve are not compatible with the fluctuations expected from the strongly simplifying assumptions of the PNM. This will be discussed in detail in Sec. 4.2. Nevertheless, it is remarkable that the agreement between model and data remains within 50% for different databases and over 9 orders of magnitude in size.

Finally, we address the question about the size (and the possible finitude) of the vocabulary of a given language as recently discussed in [MSA+11, Kle13]. Even after more than 10^6 different words the $V(N)$-data in Fig. 3.4 does not seem to saturate. To further investigate this point, we perform the PNM with the same rank-frequency distribution from Eq. (3.16) (fixed b^*, γ^*) but varying the maximum possible number of different words $V_{\mathrm{PNM}}^{\mathrm{max}}$, i.e., 1, 2, 5 ,10, and 100 times the observed number of distinct words in our largest database $Y(2000)$. It can be seen in Fig. 3.4(d) that the differences for

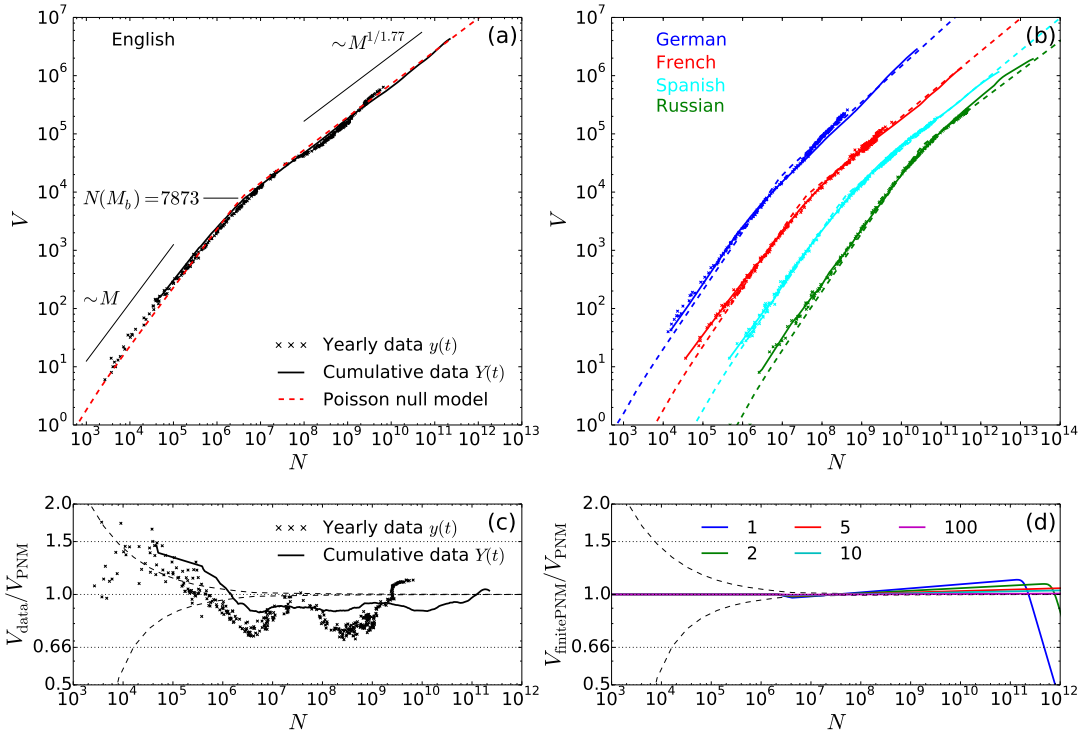

Figure 3.4.: Vocabulary V as a function of database size N (Heaps' plot). a) Number of distinct words as a function of the number of words for yearly $y(t)$ (x-symbols) database, cumulative $Y(t)$ (solid) database, and the Poisson null model (dashed) assuming $s = 41$ and the rank-frequency distribution Eq. (3.16) with $b^* = 7873$ and $\gamma^* = 1.77$. b) Same curves as in a) but for different languages showing the same scaling behaviour. In order to increase visibility the curves for French, Spanish, and Russian were shifted, respectively, by one, two, and three decades with respect to their x-values. c) Difference of the curves in a): Deviation of the data $y(t)$ and $Y(t)$ (V_{data}) from the PNM growth curve (V_{PNM}). The dashed lines show the 95%-confidence interval of the PNM. d) Deviation of a PNM growth curve with a hypothetically finite vocabulary ($V_{\mathrm{finitePNM}}$) from the PNM growth curve with infinite vocabulary (V_{PNM}) assuming rank-frequency distribution Eq. (3.16). Possible size of the total vocabulary is given in units k of the number of observed distinct words in $Y(2000)$, such that $V_{\mathrm{PNM}}^{\max} = k \cdot 4\,263\,717$ with $k = 1, 2, 5, 10, 100$. Since for $N \longrightarrow \infty : V_{\mathrm{finitePNM}}(N) \longrightarrow V_{\mathrm{PNM}}^{\max}$ the deviation for $k = 1$ becomes already large for $N > 10^{11}$.

the predicted growth curves for such different hypothetical vocabulary sizes are negligible compared to the fluctuations of the real data. From this we conclude that given the data accessible so far the possible vocabulary can be regarded for all practical purposes to be infinite (although bounded by combinatorial arguments due to a finite alphabet and word length). The fact that the same distribution Eq. (3.16) with fixed parameters accounts for the observation across all years shows that the observation of different number of words is driven mainly by the different database size and not by a change in vocabulary richness over time.

3.2.2. Preferential attachment growth model

In this section we propose a simple generative model which recovers and allows for an improved interpretation of the double scalings in our empirical findings – Eqs. (3.16) and (3.32).

Model

Our approach is different from Zipf's original explanation based on a principle of least effort between speakers and listeners [Zip36, CMFS11], but instead is in line with the tradition of preferential attachment (PA) (also known as Yule-, Simon-, or Gibrat)-type stochastic growth models explaining fat-tailed distributions [Yul25, Mit04, New05, SR10]. The main novelty in our model is that it contains two classes of word-types: a core vocabulary and a noncore vocabulary [FS01]. At each step a word (i.e. word-token) is drawn ($N \mapsto N+1$) and attributed to one of the distinct words (i.e. word-type) depending on probabilities specified below, see Fig. 3.5 for a sketch of the model.

The total number of word-types is given by $V = V_c + V_{\bar{c}}$, where ($V_{\bar{c}}$) V_c is the number of (non)core words. The new word-token can either be a new word-type ($V \mapsto V+1$) with a probability p_{new} or an already existing word-type ($V \mapsto V$) with probability $1 - p_{\text{new}}$. In the latter case, a (previously used) word-type is attributed to the word-token at random with probability proportional to the number of times this word-type has occurred before. In the former case, the new word-type can either originate from a finite set of $V_c^{\max}$ core words ($V_c \mapsto V_c + 1$) with probability p_c or come from a potentially infinite set of noncore words ($V_{\bar{c}} \mapsto V_{\bar{c}} + 1$). In our simplest model we consider p_c to be a constant, i.e. $p_c^0 \lessapprox 1$, which becomes zero only if all core words were drawn ($V_c = V_c^{\max}$):

$$p_c\left(V_c\right) = \begin{cases} p_c^0 & \text{if } V_c < V_c^{\max}, \\ 0 & \text{if } V_c = V_c^{\max}. \end{cases} \tag{3.35}$$

The final element of our model, which establishes the distinguishing aspect of core words, is the dependence of p_{new} on V. We choose p_{new} (and p_c) to depend on V and not on N because an increase in V necessarily reflects that fewer undiscovered words exist while an increase in N is strongly affected by repetitions of frequently used words. By definition, we think of core words as necessary in the creation of any text and, therefore, the usage of a new core word in a particular text should be expected and thus not affect the probability of using a new (noncore) word-type in the future, i.e., $p_{\text{new}} = p_{\text{new}}(V_{\bar{c}})$. On the other hand, if a noncore word is used for the first time ($V_{\bar{c}} \mapsto V_{\bar{c}} + 1$) the combination of this word with the previously used (core and noncore) words lead to a combinatorial increase in possibilities of expression of new ideas with the already used vocabulary and thus to a decrease in the marginal need for additional new words [PTH$^+$12]. In our model, this argument suggests that p_{new} should decrease with $V_{\bar{c}}$. Taking these factors into account, we propose as an update rule for p_{new} after each occurrence of a new noncore word as

$$p_{\text{new}} \mapsto p_{\text{new}} \left(1 - \frac{\alpha}{V_{\bar{c}} + v}\right), \tag{3.36}$$

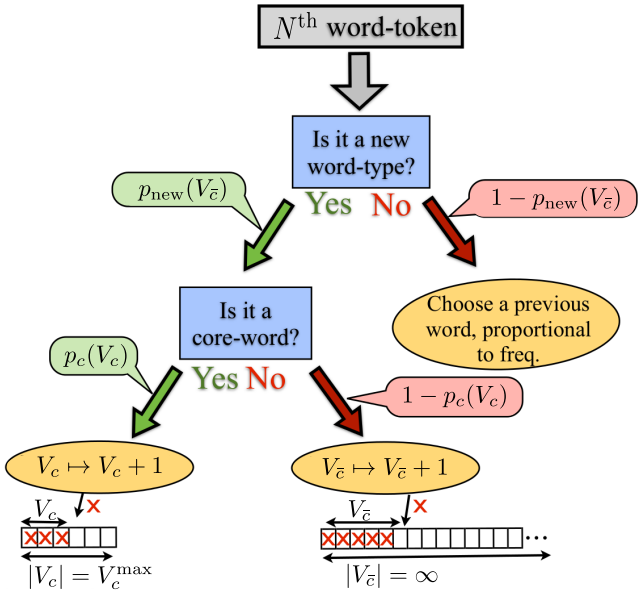

Figure 3.5.: Illustration of our generative model for the usage of new words.

with the decay rate $\alpha > 0$ and the constant $v \gg 1$ which is introduced simply in order to damp the reduction of p_{new} for small $V_{\bar{c}}$ (for simplicity, we use $v = V_c^{\max}$). The main justification for the exact functional form in Eq. (3.36) is that it allows us to recover the empirical observations reported in Figs. 3.2 and 3.4, as shown below. An alternative *a posteriori* justification will be given at the end of the next paragraph and shows that Eq. (3.36) can be interpreted as a direct consequence of an unlimited noncore vocabulary.

Analytical treatment

We now show how this model recovers Eqs. (3.16) and (3.32). We require that $1 - p_c^0 \ll 1$, which simply means that it is much more likely to draw core words than noncore words initially. In this case we can obtain approximately exact solutions for $V(N)$ in the two limiting cases considered in Eq. (3.32). When $V \ll V_c^{\max}$, which implies $V_c, V_{\bar{c}} \ll V_c^{\max}$, it follows from Eqs. (3.35) and (3.36) that $p_{\text{new}} \approx$ const. and therefore we trivially obtain that $V \sim N^1$. This case resembles the very beginning of the vocabulary growth, when most new word-types belong to the set of core words. In the case $V \gg V_c^{\max}$, $p_c = 0$ and $V \approx V_{\bar{c}}$ so that Eq. (3.36) becomes in the continuum limit:

$$\frac{\mathrm{d}}{\mathrm{d}V} p_{\text{new}}(V) = -\alpha \frac{p_{\text{new}}(V)}{V}, \tag{3.37}$$

from which it follows that $p_{\text{new}} \sim V^{-\alpha}$.

We now obtain the expected growth curve $V(N)$. Notice that our model can be considered a biased random walk in V, which, as an approximation, can be mapped onto a binomial random walk by the coordinate transformation $V(N)$ such that $p_{\text{new}}(V) = p_{\text{new}}(V(N))$. The resulting Poisson-Binomial process [Fel68] can be treated analytically, e.g., the transformation $V(N)$ is then given by the average

of the vocabulary growth:

$$\begin{aligned} V(N) &= \int_0^N \mathrm{d}N' p_{\mathrm{new}}\left(N'\right) \\ &= \int_{V(0)}^{V(N)} \mathrm{d}V' \left|\frac{\mathrm{d}N'}{\mathrm{d}V'}\right| p_{\mathrm{new}}\left(V'\right). \end{aligned} \tag{3.38}$$

Assuming $V(N) = c_1 N^{\lambda}$ the Jacobian in Eq. (3.38) gives

$$\frac{\mathrm{d}N}{\mathrm{d}V} = \frac{1}{\lambda c_1}\left(\frac{V}{c_1}\right)^{\frac{1}{\lambda}-1}. \tag{3.39}$$

Noting that $V(0) = 0$, and using $p_{\mathrm{new}} \sim V^{-\alpha}$ from above this gives for Eq. (3.38):

$$\begin{aligned} V(N) &= \int_0^{V(N)} \mathrm{d}V' \frac{1}{\lambda} c_2 c_1^{-\frac{1}{\lambda}} V'^{\frac{1}{\lambda}-\alpha-1} \\ &= \frac{1}{\lambda} c_2 c_1^{-\frac{1}{\lambda}} \frac{1}{\frac{1}{\lambda}-\alpha} V(N)^{\frac{1}{\lambda}-\alpha}. \end{aligned} \tag{3.40}$$

We, therefore, find that using $p_{\mathrm{new}} \sim V^{-\alpha}$, Eq. (3.38) holds (self-consistently) by assuming a sub-linear growth for the vocabulary $V \sim N^{\lambda}$, where the relation $\lambda = (1+\alpha)^{-1}$ is established.

In accordance with Eq. (3.32), we identify the following relation between the parameters: $V_c^{\max} = b$ and $\alpha = \gamma - 1$. The fitting parameters of Eq. (3.16) can thus be interpreted as: b is the size of the core vocabulary and γ controls the sensitivity of the probability of using a new word to the number of already used words in Eq. (3.37).

Since the probability of usage for already used word-types is assumed to be proportional to the number of times it occurred before, we guarantee that Eq. (3.32) implies (3.16) [ZM05], meaning that the double scaling in the Zipf plot is also recovered from our generative model.

Finally, we take profit of our previous calculations and provide an *a posteriori* justification of the key assumption of our model, Eq. (3.36). Our starting point is the observation – see Fig. 3.4(d) – that vocabulary is for all practical purposes infinite. We therefore postulate that

$$V(N) \overset{N\to\infty}{\longrightarrow} \infty, \tag{3.41}$$

and by following (in reverse order) the previous calculations we naturally arrive at Eq. (3.36). From the first line of Eq. (3.38) we see that in order to fulfill our postulate (3.41), p_{new} has to decay at least as slow as $p_{\mathrm{new}}(N) \sim N^{-\delta}$ with $\delta \leq 1$ for $N \to \infty$. In a minimal model it is reasonable to assume such a power law decay, in which case the first line of Eq. (3.38) implies that $V(N) \sim N^{\lambda}$ with $\lambda = 1 - \delta$. Making a transformation of variables from N to V we obtain

$$p_{\mathrm{new}}(V) = p_{\mathrm{new}}(N(V)) \sim V^{-1+\frac{1}{\lambda}} = V^{-\alpha}. \tag{3.42}$$

In turn this is equivalent to Eq. (3.37), from which we recover Eq. (3.36) as a discretized version.

Thus we see that Eq. (3.36) is a minimal assumption for an unbounded vocabulary.

Numerical simulation

In Fig. 3.6 we show direct simulations of the model in Fig. 3.5 with the traditional parameters $b = b^* = 7873$ and $\gamma = \gamma^* = 1.77$. We can clearly see that the two scaling regimes in Zipf's and Heaps' law, Eqs. (3.16, 3.32) are recovered from our model. Deviations from the data are within 50% over as much as 7 orders of magnitude. The poorer agreement for large r and N' can be attributed to a slight overestimation of the point of transition, b^*, between the two scaling regimes.

While the previous analytical arguments show that the correct scalings are obtained by our model, in order to obtain an agreement with the data it is essential to: (i) use the normalization constant C in order to determine the initial probability of finding a new word in Eq. (3.36); (ii) re-scale the distribution using the threshold $s = 41$ as $N' = N/s$; and (iii) account for the disproportionally large weight of the first word-types (in the Zipf plot), see Fig. 3.7.

In order to simulate the model, apart from fixing a number of parameters $(V_c^{\max}, \alpha, p_c^0)$, we need to prescribe how the model is initialized, e.g., what is the initial probability of using a new word p_{new}^0 and how many word-types exist at the first iteration of the model. Concerning the parameters, the initial probability of choosing a core word is set to $p_c^0 = 0.99$, such that $1 - p_c^0 \ll 1$ (see above) and the two other parameters are fixed by the fitting parameters $(V_c^{\max} = b^*, \alpha = \gamma^* - 1)$. Concerning the initialization of the model, an important point that needs to be taken into account is that we are

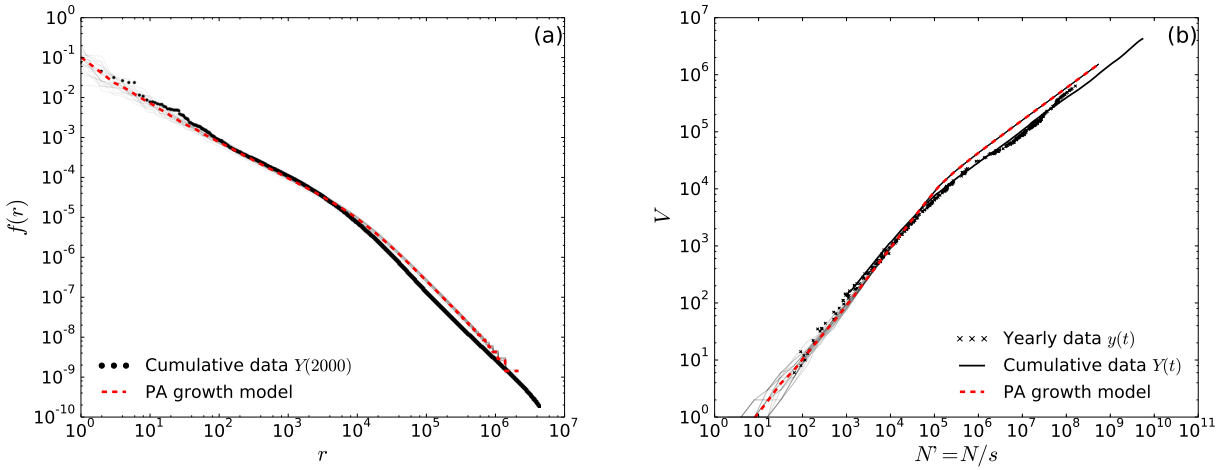

Figure 3.6.: Zipf's and Heaps' plot from the numerical simulation of our stochastic model in Fig. 3.5. (a) Rank-frequency distribution, $f(r)$, for the English database $Y(2000)$ (solid) and the expectation from our stochastic model (dashed). (b) Number of word-types as a function of word-tokens of the English database for yearly (x-symbols) database, cumulative (solid) database, and the expectation from our stochastic model (dashed). Single realizations of the stochastic process are shown in thin/gray (solid). Each realization is calculated for an imaginary text of $N' = 10^9$ tokens. Note that we exclude the word-type with rank $r = 1$ and re-normalize the remaining distribution due to the disproportional weight given to the word-types that appear first in Yule-type processes, see Fig. 3.7

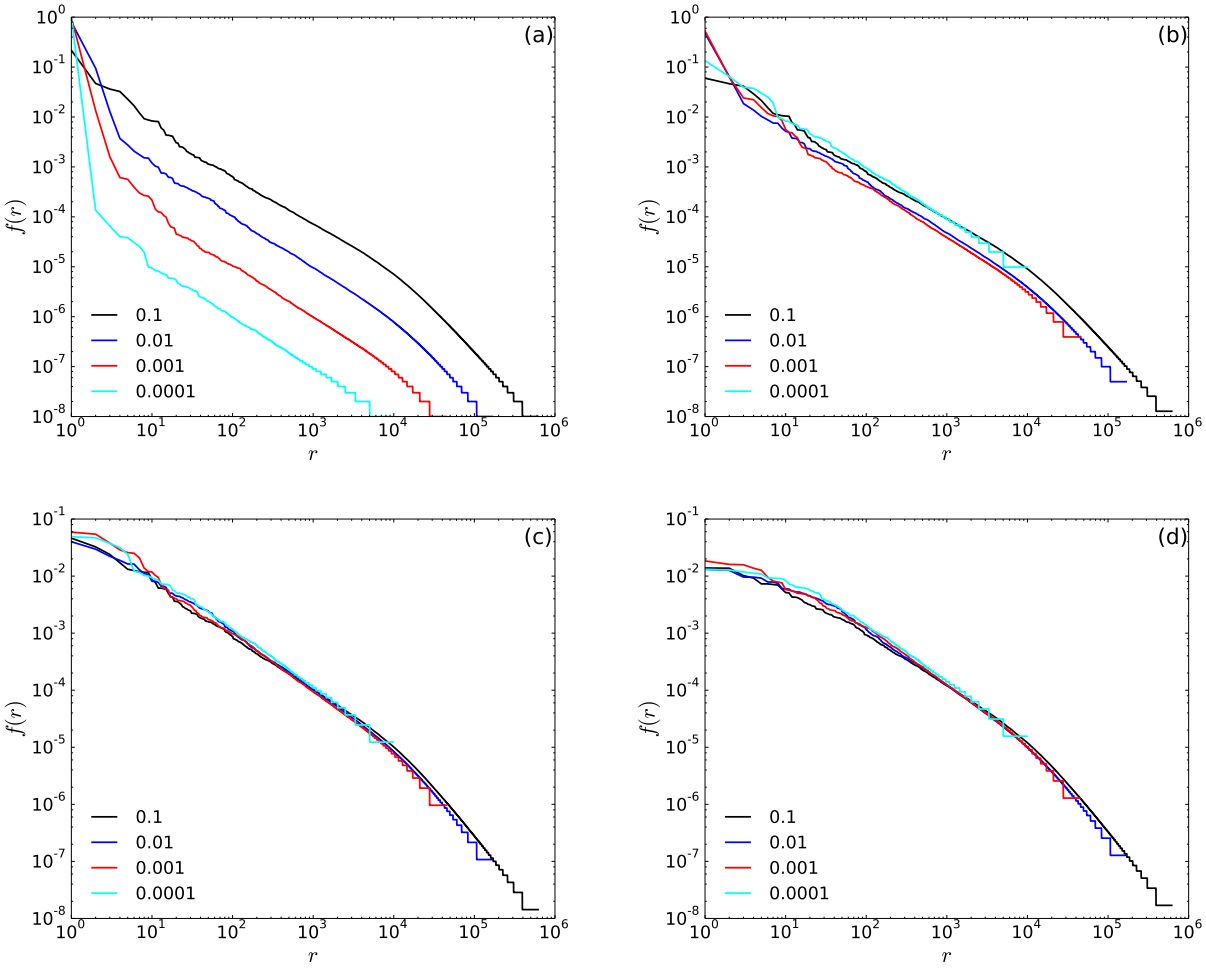

Figure 3.7.: Influence of the first word types on the rank-frequency distribution of our model. Rank-frequency distribution $f(r)$ from our numerical simulation with different values for $p^0_{\text{new}} \in \{0.1, 0.01, 0.001, 0.0001\}$ after filtering the k most frequent types, where a) $k = 0$, b) $k = 1$, c) $k = 3$, and d) $k = 10$. In this context, filtering means, that i) we neglect all tokens associated with ranks $r = 1...k$; ii) the rank of all remaining types is lowered by k, e.g., the rank of the $k + 1$-th most frequent type becomes $r = 1$; and iii) the distribution is renormalized such that $\sum_{r=1}^{V-k} f(r) = 1$, where V is the number of types before the filtering.

interested in retrieving the Heaps' plot obtained after re-scaling the number of word-tokens N by the the threshold s as $N' = N/s$ (for simplicity and computational efficiency in our simulations we choose $n = 1$). This implies that the first word-type of our model should on average appear not at the first word-token but instead approximately at $N' \approx 1/f(r = 1)$ (where $f(r = 1)$ is the frequency of the most frequent word). In view of this requirement, we set $p^0_{\text{new}} = C = f_{dp}(r = 1)$ (see Tab. 3.1) and we start with an empty list of word types (the tokens used before the appearance of the first word type are counted but not attributed to any word type). The simulations were done with a maximum number of $N = 10^9$ steps in units of word-tokens, a restriction imposed by the computational effort required. The reported results were obtained as the average of 100 realizations of the model.

Comparison to PNM

It is worth comparing the generative model with the model of random usage of words with fixed frequency, the PNM model discussed in the previous section, see Sec. 3.2.1. While the PNM allowed us to obtain Heaps' curves from Zipf's distributions (and vice-versa), in the generative model we simultaneously obtain the double scaling regime in both cases. It is important to stress that individual texts or single databases should not be considered as the output of single realizations of our generative model. Instead, we consider that not only texts but also all databases have a negligible size when compared to the language as a whole and therefore should be thought of as a small subsample ($N_{\text{database}} \ll N$) of the output of our generative model, retrieved after it achieved its stationary state ($N \to \infty$). In this case, changes in word frequencies become negligible (in the scale of N) during the creation of the database (in the sale of N_{database}). Therefore, the vocabulary growth of the created database is well approximated by the PNM with $f_{\text{dp}}(r)$.

4. Variability in word-frequency distributions

In the previous chapter we employed simple stochastic processes in order to understand the appearance of scaling laws in the statistics of word frequencies. However, the underlying assumption – that the usage of each word is governed by a Poisson process with fixed global frequency – ignores many important features of real texts and, thus, imposes severe limitations on the scope of this approach. For example, the distribution of recurrence times of individual words is characterized by non-Poissonian statistics [APM09].

In this chapter we explore this heterogeneity in the usage of words, in which we subsume any deviation from Poissonian behaviour as topicality. In this very general formulation, topical aspects in a corpus can stem from the fact that the texts were written i) by different authors; ii) in different periods in time; or iii) revolve around different topics, e.g. "sports" and "politics". The main idea in our analysis is to subdivide the corpus into smaller parts, e.g. individual chapters, individual documents, or groups of documents, and assume the validity of the Poisson assumption for each part separately. This implies that the frequency f_w of a word can not be considered as globally fixed. As a result we can study the problem of topicality by looking at the *variability of word frequencies* across the subdivided parts.

In Sec. 4.1 we a give brief overview on different measures that try to quantify topical variations on the level of individual words. In Sec. 4.2 we investigate the fluctuations around the expected vocabulary growth, finding a non-trivial scaling between the variance and the average, also known as Taylor's law. By modeling the usage of words by a simple stochastic process we show that we can account for this observation only if topical variations across different texts are considered [GA14]. In Sec. 4.3 we show how to quantify the distance between two observed texts based on the variability in the word frequencies using tools from information theory [GFCA15].

4.1. Quantifying topicality of individual words

The frequency of an individual word varies significantly across different texts meaning that its usage cannot be described alone by a single global frequency [CG95, MZ10, APM11]. For example, consider the usage of the (topical) word "network" in all articles published in the journal PlosOne. It has an overall rank $r^* = 428$ and a global frequency, $f(r^* = 428) \approx 2.9 \times 10^{-4}$, see Fig. 4.1(a). The local frequency obtained from each article separately varies over more than one decade, see Fig. 4.1(b). Note that, although in this case the local rank-ordering differs from document to document, the index r still refers to the globally determined rank and is used as a unique label for each word.

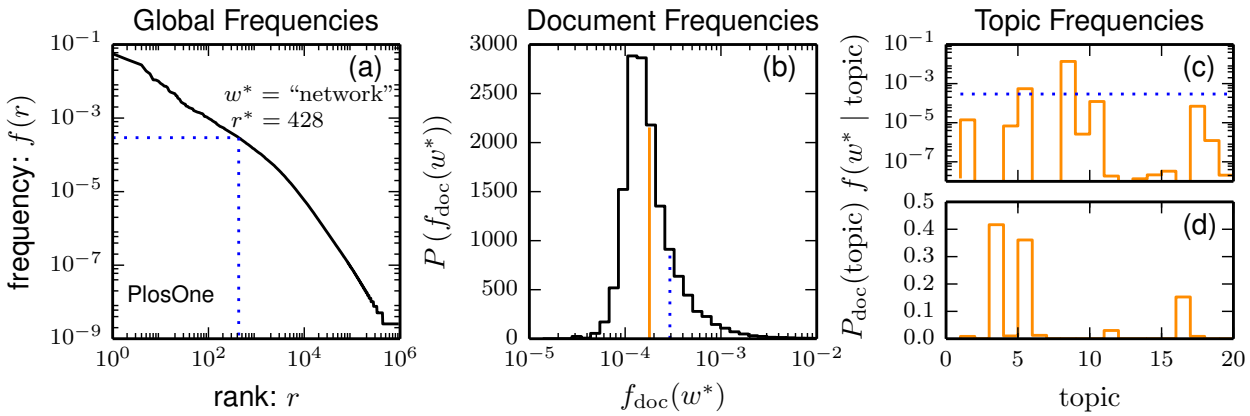

Figure 4.1.: Variation of word frequencies in the PlosOne database. (a) Rank-frequency distribution considering the complete database. The word w^* = "network" (dotted line) has $f(w^*) = f(r^* = 428) \approx 2.9 \times 10^{-4}$. (b) Distribution $P(f_{\rm doc}(w^*))$ of the local frequency $f_{\rm doc}(w^*)$ obtained from each article separately for the word "network" with the global frequency from (a) (dotted). (c) Topic-dependent frequencies $f(w^* \mid \text{topic})$ inferred from LDA with $T = 20$ topics for the word "network" with global frequency from (a) as comparison (dotted). (d) One realization for the topic composition of a single document, $P_{\rm doc}(\text{topics})$, drawn from a Dirichlet distribution. For this realization, the effective frequency is $f_{\rm doc}(w) = \sum_{t=1}^{T} P_{\rm doc}(t) f(w \mid t) \approx 2.0 \times 10^{-4}$ and is shown in (b) (solid).

Since the frequency of a word, f_w, only specifies how often a word will appear on average, the aim is to define a quantity that captures how unevenly a word is distributed among, e.g. a collection of documents. This can, for example, be applied in the automatic identification of keywords, i.e. words that carry a high informational content characterized not only by a large frequency, but also by a very uneven distribution of their occurrences.

The simplest measure is the *inverse document frequency* (IDF) [MS99]

$$IDF_w \equiv -\log \frac{D_w}{D} \tag{4.1}$$

which counts the number of documents D_w (out of D documents in total) where word w appears at least once. Since IDF_w is highly correlated to the frequency f_w, one can calculate the expected $IDF_w^{(0)}$ from a Poisson null model in which one assumes that i) the word is used with f_w in each document; and ii) each document is of the same length $n = \frac{1}{D}\sum_d n_d$ (where n_d is the number of word-tokens of document d) giving

$$IDF_w^{(0)} = -\log \frac{D(1 - e^{-f_w n})}{D} = -\log(1 - e^{-f_w n}) \tag{4.2}$$

From this one defines the *residual inverse document frequency* ($rIDF$) [CG95] as the difference between the measured and the expected IDF as

$$rIDF_w \equiv IDF_w - IDF_w^{(0)} = -\log \frac{D_w}{D(1 - e^{-f_w n})}. \tag{4.3}$$

Accounting for the unequal length of documents in Eq. (4.2), Ref. [APM11] introduced a generalization of $rIDF$ called *dissemination*, U, defined as

$$U_w \equiv \frac{D_w}{\sum_{d=1}^{D}(1 - e^{-f_w n_d})}. \tag{4.4}$$

We can immediately see that for the case of equal document lengths, i.e. $n_d = n$, we get that $rIDF_w = -\log U_w$.

A different and even more general approach in the framework of information theory was proposed in Ref. [MZ10], where the authors calculate the *information content* h_w of a single word w. The starting point is the conditional probability $p(d \mid w)$, i.e. the probability that a word-token belongs to document d given that we know that its associated word-type is w,

$$p(d \mid w) = p(w \mid d)\frac{p(d)}{p(w)} = n_{w,d}/n_d \frac{n_d/N}{n_w/N} = \frac{n_{w,d}}{n_w}, \tag{4.5}$$

where $n_{w,d}$ is the number of times word w appears in document d and $n_w = \sum_d n_{w,d}$ is the total number of occurrences of word w. Noting that $p(d \mid w)$ is normalized, i.e. $\sum_d p(d \mid w) = 1$, one can calculate an entropy-like quantity

$$H(\{d\} \mid w) \equiv -\sum_{d=1}^{D} p(d \mid w) \log p(d \mid w). \tag{4.6}$$

The information content h_w is then defined as the difference between the measured $H(\{d\} \mid w)$ and $H^{(0)}(\{d\} \mid w)$, where all word-tokens are shuffled across all different texts as

$$h_w \equiv H^{(0)}(\{d\} \mid w) - H(\{d\} \mid w). \tag{4.7}$$

Here, we show how this measure can be related to the IDF, Eq. (4.1), by assuming that for a word w, we only know D_w (the number of documents it appears at least once), such that we can approximate $p(d \mid w)$ in $H(\{d\} \mid w)$ as

$$p(d \mid w) \approx \begin{cases} 1/D_w, & d = 1, \ldots, D_w \\ 0, & d = D_w + 1, \ldots, D \end{cases} \tag{4.8}$$

which gives $H(\{d\} \mid w) \approx \log D_w$. Likewise, for the shuffled version of $H^{(0)}(\{d\} \mid w)$ we approximate $p(d \mid w) \approx 1/D$ for $d = 1, \ldots, D$ such that we get $H^{(0)}(\{d\} \mid w) = \log D$ which yields

$$h_w \approx -\log \frac{D_w}{D} = IDF_w. \tag{4.9}$$

This demonstrates that the previously (ad-hoc) defined measures in Eqs. (4.1), (4.3), and (4.4) can be retrieved as special cases of a more general formulation, Eq. (4.7), offering an information-theoretic interpretation of the respective quantities.

Note that the authors in Ref. [MZ10] also consider the integrated (over all words w) quantity

$$MI(\{d\};\{w\}) \equiv \sum_w f_w h_w \tag{4.10}$$

which they identify as the mutual information between the distribution of words w and the partitioning into documents d [Zan14], i.e. it quantifies the average amount of information one obtains from a randomly sampled word-token about which of the documents it belongs to. In Ref. [GBGC+02] it was shown that this measure corresponds to the Jensen-Shannon divergence, which will be treated in detail in Sec. 4.3.

Another popular approach to account for the heterogeneity in the usage of single words are *topic models* [Ble12]. The basic idea is that the variability across different documents can be explained by the existence of (a smaller number of) topics. In the framework of a generative model it assumes i) that individual documents are composed of a mixture of topics (indexed by $t = 1, .., T$), with each topic represented in an individual document by the probabilities $P_{\rm doc}(\text{topic} = t)$; and ii) that the frequency of each word is topic-dependent, i.e. $f(r \mid \text{topic} = t)$, which leads to a different effective frequency in each document, $f_{\rm doc}(r) = \sum_{t=1}^{T} P_{\rm doc}(t) f(r \mid t)$. One particularly popular variant of topic models is Latent Dirichlet Allocation (LDA) [BNJ03], which assumes that the topic composition $P_{\rm doc}(\text{topic})$ of each document is drawn from a Dirichlet distribution, $P_{\rm Dir}$, such that only few topics contribute substantially to each document. Given a database of documents, LDA infers the topic-dependent frequencies, $f(r \mid \text{topic})$, from numerical maximization of the posterior likelihood of the generative model [ŘS10]. As an illustration, in Fig. 4.1(c) we show $f(r^* \mid \text{topic})$ obtained using LDA for the word "network" in the PlosOne database. As expected from a meaningful topic model, we see that the conditional frequencies vary over many orders of magnitude, and that the global frequency $f(r^*)$ is governed by few topics. The advantage of LDA is that, instead of measuring the distribution of frequencies of each individual word (or 2-point distributions for assessing correlations) over different documents, it estimates the frequency of individual words for a finite (and small) number of topics. In combination with the generative model (e.g., drawing $P_{\rm doc}(\text{topic})$ from a Dirichlet distribution), this not only yields a more compact description of topicality by dramatically reducing the number of parameters, but also allows for an easy extrapolation to unseen texts from a small training sample [BNJ03]. Note that, here, we simply wish to illustrate the concept of topicality by means of non-global word frequencies, in which topic models are one useful approach. The idea of topic models will be discussed in detail in Ch. 5.

4.2. Fluctuations in the vocabulary growth: Taylor's law

In this section, we consider the problem of the vocabulary growth for an ensemble of texts investigating the fluctuations around the Heaps' law. We study the scaling of fluctuations by looking at the relation between the standard deviation, $\sigma(N) = \sqrt{\mathbb{V}[V(N)]}$, and the mean value, $\mu(N) = \mathbb{E}[V(N)]$, computed over the ensemble of texts with the same text length N. In other systems, Taylor's law [Tay61]

$$\sigma(N) \propto \mu(N)^{\beta} \text{ for } \mu(N) \gg 1 \tag{4.11}$$

with $1/2 \leq \beta \leq 1$ is typically observed [EBK08].

Looking at the expected fluctuations (e.g., for Heaps' law) quantitatively is important, e.g., when one wants to test the validity of the law for actual observed data as argued in Sec. 3.1. The scaling behaviour, Eq. (4.11), in the form of the exponent β describes the self-averaging property of the analyzed variable (here: the vocabulary growth) in the form of the normalized variance $\tilde{\sigma}(N) = \sigma(N)/\mu(N) = \mu(N)^{\beta-1}$ [Sor06]. If $\tilde{\sigma}(N) \to 0$ in the limit $N \to \infty$, the vocabulary growth is self-averaging. This implies that the parameter β determines what can be considered a sufficiently large sample size N such that a single realization is representative of the whole ensemble. While for uncorrelated samples the convergence due to the central limit theorem leads to $\beta = 1/2$ [EBK08], Taylor's law with $1/2 < \beta < 1$ found in many social systems implies a much slower convergence of the relative fluctuations with sample size. For exponents $\beta \geq 1$, even in the limit $N \to \infty$ the relative fluctuations are still finite, i.e. $\lim_{N\to\infty} \tilde{\sigma}(N) \to c > 0$; in this case the variable is called nonself-averaging.

In Sec. 4.2.1 we present empirical evidence for Taylor's law with $\beta = 1$ in the vocabulary growth in written text with focus on the deviations from the Poisson null model from Sec. 3.2.1. In Sec. 4.2.2 we show how these deviations can be explained by accounting for topical aspects of written text. In an extension of the Poisson null model, this aspect plays the role of a quenched disorder and leads to a nonself-averaging process. The consequences of our findings to applications, e.g. vocabulary richness, are discussed in Sec. 4.2.3.

4.2.1. Empirical evidence

In Fig. 4.2 we show empirical data of real texts from three databases (Wikipedia, PlosOne, and Google-ngram, see Appendix A.1-A.3 for details on the data) for the scaling relations of Zipf's, Heaps', and Taylor's law, Eqs. (3.16,3.32,4.11), and compare them with predictions from the Poisson null model in Eqs. (3.19,3.20). The Poisson null model correctly elucidates the connection between the scaling exponents in Zipf's and Heaps' law as argued in Sec. 3.2.1, but it suffers from two severe drawbacks. First, it is of limited use for a quantitative prediction of the vocabulary size for individual articles as it systematically overestimates its magnitude, see Fig. 4.2(b,e,h). Second, it dramatically underestimates the expected fluctuations of the vocabulary size yielding a qualitatively different behavior in the fluctuation scaling: whereas the Poisson null model yields an exponent $\beta \approx 1/2$ expected from central-limit-theorem-like convergence [EBK08], the three empirical data [Fig. 4.2(c,f,i)] exhibit a scaling with $\beta \approx 1$. This implies that relative fluctuations of V around its mean value μ for fixed N do not decrease with larger text size (the vocabulary growth, $V(N)$, is a nonself-averaging quantity) and remain of the order of the expected value. Indeed, we find that in all three databases

$$\sigma(N) \approx 0.1\mu(N). \tag{4.12}$$

Instead of looking at a single value (V, N) for each document, as described above, an alternative approach is to count the number of different words, V, in the first N words of the document. This leads to a curve $V(N)$ for $N = 1, 2, \ldots, N_{\max}$, where $N_{\max}$ is the length of the document. This alternative approach was employed in Fig. 4.2(e,f) and leads to results equivalent to the ones obtained using single values (V, N), i.e. the $\mu(N)$ and $\sigma(N)$ obtained over different texts lead to identical Heaps' and Taylor's laws. In Fig. 4.2(f) we show that anomalous fluctuation scaling in the vocabulary growth is preserved if shuffling the word order of individual texts. This illustrates that in contrast to usual explanations of fluctuation scaling in terms of long-range correlations in time series [EBK08], here, the observed deviations from the Poisson null model are mainly due to fluctuations across different texts.

In the following, we argue that these observations can be accounted for by considering topical aspects of written language, i.e. instead of treating word frequencies as fixed, we will consider them to be topic-dependent ($f(r) \mapsto f(r \mid \text{topic})$).

4.2.2. Vocabulary growth with variable word frequencies

Extended Poisson null model

In this section we show how topicality can be included in the analysis of the vocabulary growth. The simplest approach is to consider again that the usage of each word is governed by Poisson processes, but this time to consider that frequencies are not fixed but are themselves random variables that vary across texts.

In this setting, the random variable representing the vocabulary size, V, for a text of length N can be written as

$$V(N)^{(i,j)} = \sum_r I\left[n_r^{(i)}(N, f^{(j)}(r))\right], \tag{4.13}$$

in which n_r is the integer number of times the word r occurs in a Poisson process of length N with frequency $f(r)$ and $I[x]$ is an indicator-type function, i.e. $I[x = 0] = 0$ and $I[x \geq 1] = 1$. Here, we introduced two additional indices: i) the index i characterizes a realization of the Poisson processes $n_r^{(i)}(N, f^{(j)}(r))$ for a given realization j of the set of frequencies $f^{(j)}(r)$; and ii) the index j characterizes a realization of the set of frequencies $f^{(j)}(r)$ (which vary due to topicality). The corresponding averages, therefore, not only consist of averaging over the realizations of the Poisson process (indexed by i) as presented in Sec. 3.2.1, but additionally require the averaging over all possible realizations of the sets of frequencies (indexed by j). In this framework expectation values correspond to quenched averages which we will denote by subscript q, i.e. $\mathbb{E}_q[X] = \left\langle X^{(i,j)} \right\rangle_{i,j}$. In the following we are interested in $\mathbb{E}_q\left[V(N)\right]$ and $\mathbb{E}_q\left[V(N)^2\right]$ in order to obtain results of the average and variance in analogy to Eqs. (3.19, 3.20).

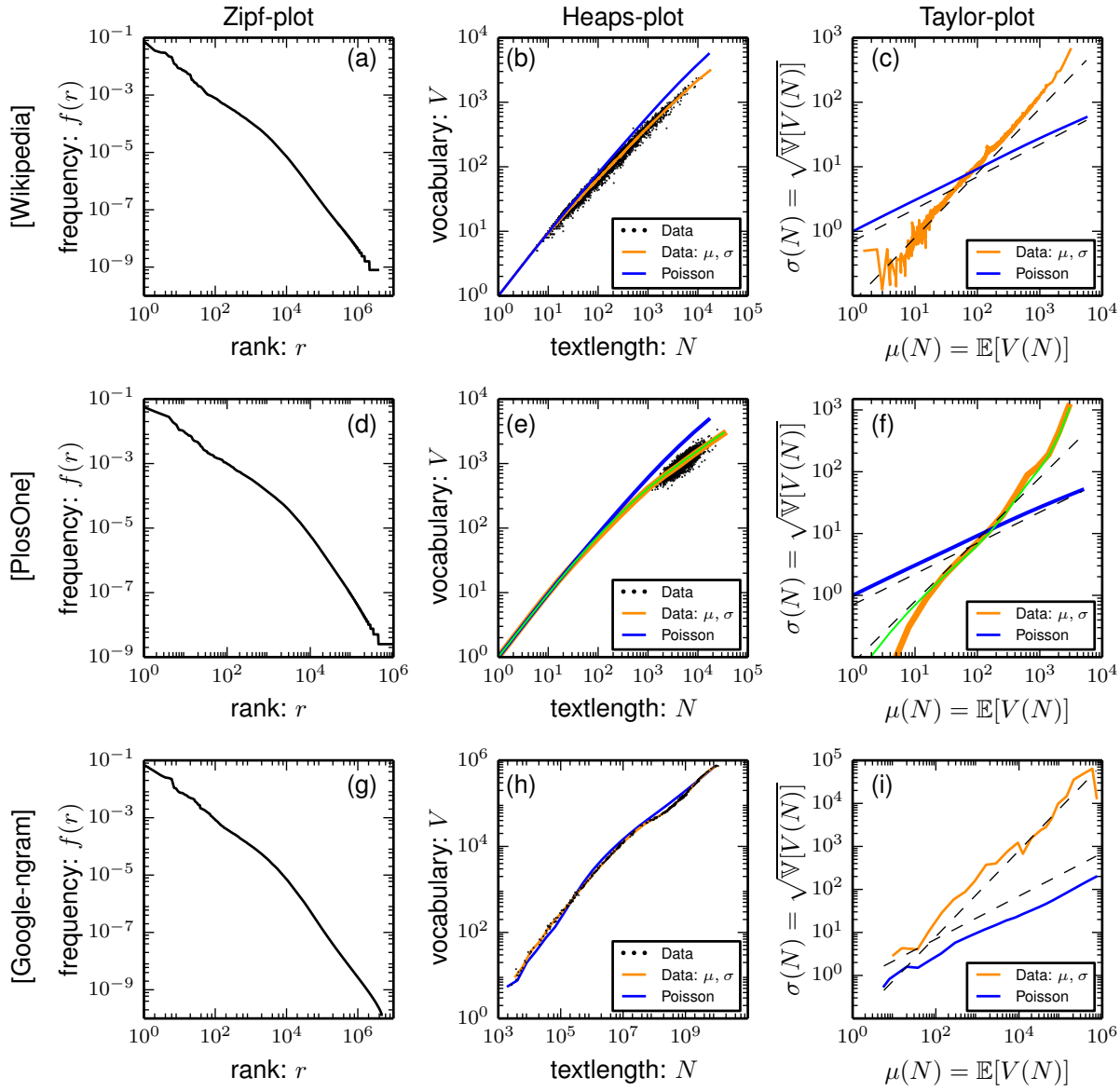

Figure 4.2.: Scaling of Zipf's law (3.2), Heaps' law (3.18), and Taylor's law (4.11). Each row corresponds to one of the three databases used in this analysis. (a,d,g) Zipf's law: Rank-frequency distribution $f(r)$ considering the full database. (b,e,h) Heaps' law: the number of different words, V, as a function of textlength, N, for each individual article in the corresponding database (black dots). (c,f,i) Taylor's law: standard deviation, $\sigma(N)$, as a function of the mean, $\mu(N)$, for the vocabulary $V(N)$ conditioned on the textlength N. Poisson (blue-solid) shows the expectation from the Poisson null model, Eqs. (3.19,3.20), assuming the empirical rank-frequency distribution from (a,d,g), respectively. (Data: μ, σ) (yellow-solid) shows the mean, $\mu(N)$, and standard deviation, $\sigma(N)$, of the data $V(N)$ within a running window in N. Additionally, (e,f) show the results (Data: μ, σ) obtained shuffling the word order for each individual article (thin green-solid). The fact that this curve is indistinguishable from the original curve shows that the results are not due to temporal correlations within the text. For comparison, we show in (c,f,i) the scalings $\sigma(N) \propto \mu(N)^{1/2}$ and $\sigma(N) \propto \mu(N)$ (dashed).

We start by noting that for a fixed realization j our model is equivalent to the PNM, Sec. 3.2.1

$$\left\langle I\left[n_r^{(i)}(N, f^{(j)}(r))\right]\right\rangle_j = 1 - P(n_r = 0; N, f^{(j)}(r)) = 1 - e^{-Nf^{(j)}(r)}. \tag{4.14}$$

which corresponds to the probability of word r not occurring for a Poisson process of duration N with frequency $f^{(j)}(r)$. Therefore, we get for $\mathbb{E}_q[V(N)]$:

$$\mathbb{E}_q[V(N)] = \left\langle V(N)^{(i,j)}\right\rangle_{i,j} = \sum_r \left\langle I\left[n_r^{(i)}(N, f^{(j)}(r))\right]\right\rangle_{i,j} = \sum_r 1 - \langle e^{-Nf^{(j)}(r)}\rangle_j, \tag{4.15}$$

Using that $I[x]^2 = I[x]$, and that two Poisson processes of different words ($r \neq r'$) with a given set of frequencies $f^{(j)}(r)$ are independent of each other, we obtain for $\mathbb{E}_q\left[V(N)^2\right]$

$$\begin{aligned}
\mathbb{E}_q\left[V(N)^2\right] &= \left\langle V(N)^{(i,j)}V(N)^{(i,j)}\right\rangle_{i,j} \\
&= \left\langle \sum_{r,r'} I[n_r^{(i)}(N, f^{(j)}(r))]I[n_{r'}^{(i)}(M, f^{(j)}(r'))]\right\rangle_{i,j} \\
&= \sum_r \left\langle I[n_r^{(i)}(N, f^{(j)}(r))]^2\right\rangle_{i,j} + \left\langle \sum_r \sum_{r'\neq r} I[n_r^{(i)}(N, f^{(j)}(r))]I[n_{r'}^{(i)}(N, f^{(j)}(r'))]\right\rangle_{i,j} \\
&= \sum_r \left\langle I[n_r^{(i)}(N, f^{(j)}(r))]\right\rangle_{i,j} + \sum_r \sum_{r'\neq r} \left\langle \langle I[n_r^{(i)}(N, f^{(j)}(r))]\rangle_i \langle I[n_{r'}^{(i)}(M, f^{(j)}(r'))]\rangle_i\right\rangle_j \\
&= \sum_r 1 - \left\langle e^{-Nf(r)}\right\rangle_j + \sum_r \sum_{r'\neq r} \left\langle \left(1 - e^{-Nf(r)}\right)\left(1 - e^{-Nf(r')}\right)\right\rangle_j
\end{aligned} \tag{4.16}$$

For simplicity, hereafter $\langle \ldots \rangle \equiv \langle \ldots \rangle_j$ (the average over realizations of sets of frequencies $f^{(j)}(r)$).

Using the inequality between arithmetic and geometric mean

$$e^{\langle \ln x\rangle} = \langle x\rangle_{\text{geometric}} \leq \langle x\rangle_{\text{arithmetic}} = \left\langle e^{\ln x}\right\rangle, \tag{4.17}$$

we obtain that

$$\mathbb{E}_q[V(N)] = \sum_r 1 - \left\langle e^{-Nf(r)}\right\rangle \leq \sum_r 1 - e^{-N\langle f(r)\rangle} \equiv \mathbb{E}_a[V(N)]. \tag{4.18}$$

The right hand side corresponds to the result of the Poisson null model (with fixed $f(r) = \langle f(r)\rangle$), see Eq. (3.19), and can be interpreted as an annealed average (denoted by subscript a). This implies that the heterogeneous dissemination of words across different texts leads to a reduction of the expected size of the vocabulary, in agreement with the first deviation of the Poisson null model reported in Fig. 4.2(b,e,h).

For the quenched variance we obtain

$$\begin{aligned}\mathbb{V}_q\left[V(N)\right] \equiv& \mathbb{E}_q\left[V(N)^2\right] - \mathbb{E}_q\left[V(N)\right]^2 \\ =& \sum_r \left\langle e^{-Nf(r)} \right\rangle - \left\langle e^{-Nf(r)} \right\rangle^2 + \sum_r \sum_{r' \neq r} \mathrm{Cov}[e^{-Nf(r)}, e^{-Nf(r')}]\end{aligned} \tag{4.19}$$

where $\mathrm{Cov}[e^{-Nf(r)}, e^{-Nf(r')}] \equiv \left\langle e^{-Nf(r)} e^{-Nf(r')} \right\rangle - \left\langle e^{-Nf(r)} \right\rangle \left\langle e^{-Nf(r')} \right\rangle$. Comparing to the Poisson case in Eq. (3.20), we see that the quenched average yields an additional term containing the correlations of different words. In general, this term does not vanish due to the semantic relation between words (i.e. some words are more likely to occur in the presence of other words, e.g. "quantum" and "physics") and is responsible for the anomalous fluctuation scaling with $\beta = 1$ observed in real text, explaining the second deviation from the Poisson null model reported in Fig. 4.2(c,f,i).

Specific ensembles

In this section we compute the general results from Eqs. (4.15,4.19) for particular ensembles of frequencies $f^{(j)}(r)$ and compare them to the empirical results. In the absence of a generally accepted parametric formulation of such an ensemble, we propose two nonparametric approaches explained in the following.

In the first approach we construct the ensemble $f^{(j)}(r)$ directly from the collection of documents, i.e. the frequency $f^{(j)}(r)$ corresponds to the frequency of word r in document j, such that

$$\left\langle e^{-Nf(r)} \right\rangle = \frac{1}{D} \sum_{j=1}^{D} e^{-Nf^{(j)}(r)}, \tag{4.20}$$

where D is the number of documents in the data, see Fig. 4.1(b).

In the second approach we construct the ensemble from the LDA topic model [BNJ03], in which $f^{(j)}(r) = f(r \mid \mathrm{topic} = j)$ corresponds to the frequency of word r conditional on the topic $j = 1...T$, see Fig. 4.1(c+d). In this particular formulation each document is assumed to consist of a composition of topics, $P_{\mathrm{doc}}(\mathrm{topic})$, which is drawn from a Dirichlet distribution, such that we get for the quenched average

$$\left\langle e^{-Nf(r)} \right\rangle = \int \mathrm{d}\theta P_{\mathrm{Dir}}(\theta|\alpha) e^{-Nf(r;\theta)}, \tag{4.21}$$

in which $\theta = (\theta_1, ..., \theta_T)$ are the probabilities of each topic, $f(r;\theta) = \sum_{j=1}^{T} \theta_j f(r \mid \mathrm{topic} = j)$, and the integral is over a T-dimensional Dirichlet distribution $P_{\mathrm{Dir}}(\theta|\alpha)$ with concentration parameter $\alpha = 1.0$. We infer the $f(r \mid \mathrm{topic})$ using Gensim [ŘS10] for LDA with $T = 100$ topics.

The results from both approaches are compared to the PlosOne database in Fig. 4.3. Fig. 4.3(a) shows that both methods lead to a reduction in the mean number of different words. Whereas the direct ensemble, Eq. (4.20), almost perfectly matches the curve of the data, the LDA-ensemble, Eq. (4.21), still overestimates the mean number of different words in the data. This is not surprising

since due to the fewer number of topics (when compared to the number of documents) it constitutes a much more coarse-grained description than the direct ensemble. Additionally, the LDA-ensemble relies on a number of ad-hoc assumptions, e.g. the Dirichlet distribution in Eq. (4.21) or the particular choice of parameters in the inference algorithm which were not optimized here. More importantly, both methods correctly account for the anomalous fluctuation scaling with $\beta = 1$ observed in the real data, see Fig. 4.3(b) and even yield a similar proportionality factor in the quantitative agreement with the data. The comparison of the individual contributions to the fluctuations, Eq. (4.19), shown in the inset of Fig. 4.3(b) shows that the anomalous fluctuation scaling is due to correlations in the co-occurrence of different words (contained in the term $\mathrm{Cov}[e^{-Nf(r)}, e^{-Nf(r')}]$).

Example: Vocabulary growth for Gamma-distributed frequency and a double power law

In this section, we provide a simple example in which we can obtain a closed-form expression for the quenched average of the vocabulary growth, Eq. (4.15), which we will use in Sec. 4.2.3.

For this, we i) assume that the average rank-frequency distribution, $\langle f(r)\rangle$ is given by the double power law we found in Sec. 3.1, Eq. (3.16); and ii) follow the suggestion by [CG95] that the distribution of the frequency of single words across different texts can be roughly described by a Gamma

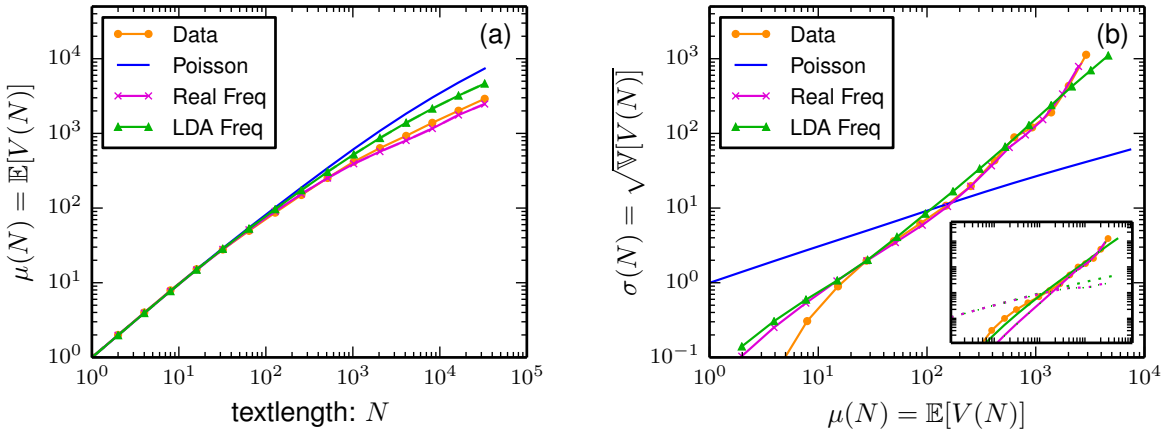

Figure 4.3.: Vocabulary growth for specific topic models. (a) Average vocabulary growth and (b) fluctuation scaling in the PlosOne database (Data) and in the calculations from Eqs. (4.15,4.19) for the two topic models based on the measured frequencies in individual articles (Real Freq) and on LDA (LDA Freq), compare Eqs. (4.20,4.21). For comparison we show the results from the Poisson null model (Poisson), Eqs. (3.19,3.20), which do not take into account topicality. The inset in (b) (same scale as main figure) shows the individual contributions to the fluctuations in Eq. (4.19): $\sum_r \left\langle e^{-Nf(r)}\right\rangle - \left\langle e^{-2Nf(r)}\right\rangle$ (dotted) and $\sum_r \sum_{r'\neq r} \mathrm{Cov}[e^{-Nf(r)}, e^{-Nf(r')}]$ (solid), illustrating that correlations between different words lead to anomalous fluctuation scaling. The solid lines for LDA-Freq and Real Freq in (b) show the calculations of the corresponding topic models replacing the Poisson by multinomial usage in the derivation of Eqs. (4.15,4.19) in order to avoid finite-size effects for $\mu(N) < 100$.

distribution:

$$P_\Gamma(f(r)=x;a,\eta)=\frac{1}{\Gamma(a)}\eta^{-a}x^{a-1}e^{-x/\eta} \tag{4.22}$$

such that we can calculate the quenched average

$$\left\langle e^{-Nf(r)}\right\rangle=\int \mathrm{d}x P_\Gamma(f(r)=x;a,\eta)e^{-Nx}=(1+\eta N)^{-a}. \tag{4.23}$$

If we assume that the distribution of frequencies for all words is given by the same shape-parameter a (e.g. $a=1$ corresponds to an exponential distribution) and fix the mean of the distribution, given by $\langle f(r)\rangle = a\eta$ we get $\left\langle e^{-Nf(r)}\right\rangle = (1+N\langle f(r)\rangle/a)^{-a}$.

From the assumption of a double power law, Eq. (3.16), for $\langle f(r)\rangle = f_{\mathrm{dp}}(r;\gamma,b)$ we can calculate the vocabulary growth according to Eq. (3.19) analytically in the continuum approximation by substituting $x\equiv\langle f(r)\rangle$:

$$\begin{aligned}\mathbb{E}_q\left[V(N)\right] &= \sum_r 1-(1+N\langle f(r)\rangle/a)^{-a} &(4.24)\\ &= -\int_0^1 \mathrm{d}x\frac{\mathrm{d}r}{\mathrm{d}x}\left[1-(1+Nx/a)^{-a}\right] &(4.25)\end{aligned}$$

which can be expressed in terms of the ordinary hypergeometric function $H\equiv {}_2F_1$ [AS72] yielding

$$\begin{aligned}\mathbb{E}_q\left[V(N)\right] &= b-C+b\left[H(a,-\frac{1}{\gamma},1-\frac{1}{\gamma},-\frac{CN}{ab})-1\right]\\ &- C\left(1+\frac{N}{a}\right)^{-a}\left[a\frac{\Gamma(1+a)}{\Gamma(2+a)}H(1,1,2+a,-\frac{a}{N})-1\right]\\ &+ b\left(1+\frac{CN}{ab}\right)^{-a}\left[a\frac{\Gamma(1+a)}{\Gamma(2+a)}H(1,1,2+a,-\frac{ab}{CN})-1\right],\end{aligned} \tag{4.26}$$

where the vocabulary growth $\mathbb{E}_q\left[V(N)\right]$ is parametrized by γ, b, and a (note that the parameter $C=C(\gamma,b)$ is the normalization constant in $f_{\mathrm{dp}}(r;\gamma,b)$, see Tab. 3.1).

In the limit $a\to\infty$ the Gamma distribution $P_\Gamma(f(r)=x;a,\eta)$ with given mean $\langle f(r)\rangle = a\eta =$ const. converges to a Gaussian with $\sigma^2=\langle f(r)\rangle^2/a$. For $a\to\infty$, $\sigma^2\to 0$ and we recover the Poisson null model, Eqs. (3.19,3.20), in which the individual frequencies $f(r)$ are fixed (annealed average).

4.2.3. Application: Measuring vocabulary richness

When measuring vocabulary richness we want a measure which is robust to different text sizes. The traditional approach is to use Herdan's C, i.e. $C_H=\log V/\log N$ [WA99, Baa01, YKK12]. While quite effective for rough estimations, this approach has several problems. One obvious one is that it does not incorporate any deviations from the original Heaps' law (e.g., the double scaling regime in Sec. 3.1). More seriously, it does not provide any estimation of the statistical significance or expected fluctuations of the measure. For instance, if two values are measured for different texts one can not

determine whether one is significantly larger than the other. Our approach is to compare observations with the fluctuations expected from models in the spirit of the extended PNM, Eqs. (4.15,4.19).

The computation of statistical significance requires an estimation of the probability of finding V different words in a text of length N, $P(V|N)$, which can be obtained from a given generative model, e.g. Eqs. (4.15,4.19). For a text with (V^*, N^*) we compute the percentile $P(V > V^*|N^*)$, which allows for a ranking of texts with different sizes such that the smaller the percentile, the richer the vocabulary. An estimation of the significance of the difference in the vocabulary can then be obtained by comparison of the different percentile.

For the sake of simplicity, we illustrate this general approach by approximating $P(V|N)$ by a Gaussian distribution. In this case, the percentile are determined by the mean, $\mu(N) = \mathbb{E}[V(N)]$, and the variance, $\sigma(N) = \sqrt{\mathbb{V}[V(N)]}$, in terms of the z-score

$$z_{(V,N)} = \frac{V - \mu(N)}{\sigma(N)}, \tag{4.27}$$

which shows how much the measured value (V, N) deviates from the expected value $\mu(N)$ in units of standard deviations ($z_{(V,N)}$ follows a standard normal distribution: $z \overset{d}{\sim} \mathcal{N}(0,1)$). If we take into account our quantitative result on fluctuation scaling in the vocabulary in Eq. (4.12), i.e. $\sigma(N) \approx 0.1\mu(N)$, we can calculate the z-score of the observation (V, N) as

$$z_{(V,N)} \approx \frac{V - \mu(N)}{0.1\mu(N)} = 10\left(\frac{V}{\mu(N)} - 1\right), \tag{4.28}$$

in which we need to include the expected vocabulary growth, $\mu(N)$, from a given generative model (e.g., Heaps' law with two scalings). We can now: i) for a single text (V, N), assign a value of lexical richness, the z-score $z_{(V,N)}$, taking into account deviations from the pure Heaps' law which should be included in $\mu(N)$; ii) given two texts (V_1, N_1) and (V_2, N_2), compare directly the respective z-scores $z_{(V_1,N_1)}$ and $z_{(V_2,N_2)}$ in order to assess which text has a higher lexical richness independent of the difference in the text lengths; and iii) estimate the statistical significance of the difference in vocabulary by considering $\Delta z \equiv z_{(V_1,N_1)} - z_{(V_2,N_2)}$, which is distributed according to $\Delta z \overset{d}{\sim} \mathcal{N}(0,2)$ since $z \overset{d}{\sim} \mathcal{N}(0,1)$. Point (iii) implies that the difference in the vocabulary richness of two texts is statistically significant on a 95%-confidence level if $|\Delta z| > 2.77$, i.e. in this case there is at most a 5% chance that the observed difference originates from topic fluctuations. As a rule of thumb, for two texts of approximately the same length ($V(N) \approx \mu(N)$), the relative difference in the vocabulary has to be larger than 27.7% in order to be sure on a 95%-confidence level that the difference is not due to expected topic fluctuations.

We illustrate this approach for the vocabulary richness of Wikipedia articles. As a proxy for the true vocabulary richness, we measure how much the vocabulary of each article, $V(N)$, exceeds the average vocabulary $V_{\text{avg}}(N)$ with the same text length N empirically determined from all articles in the Wikipedia. In practice, however, when assessing the vocabulary richness of a single article, information of $V_{\text{avg}}(N)$ from an ensemble of texts is usually not available and measures such as the

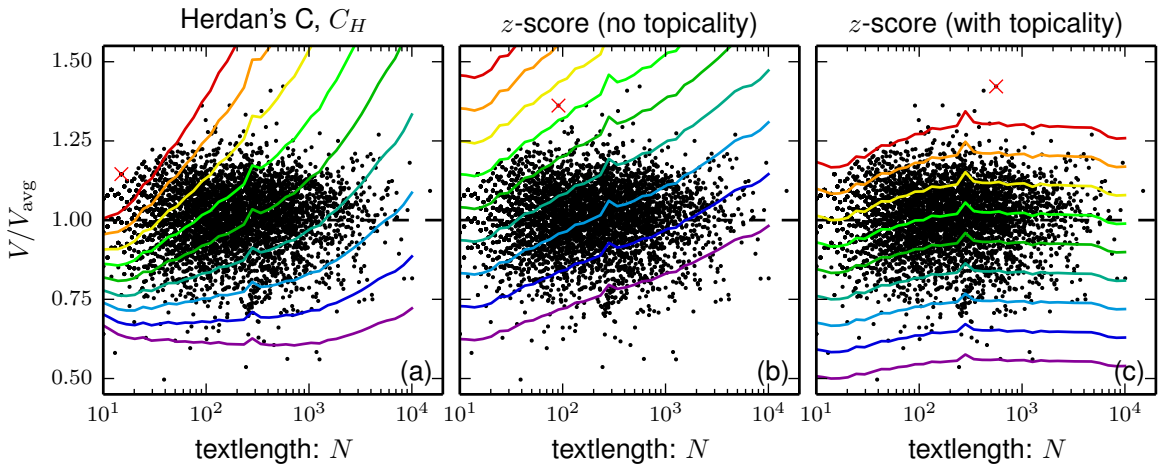

Figure 4.4.: Measures of vocabulary richness. For 5000 randomly selected articles from the Wikipedia database (black dots), we compute the ratio between the number of different words $V(N)$ and the average number of different words $V_{\rm avg}(N)$ (empirically determined from all articles with the same textlength N). We compare the predictions of different measures of vocabulary richness (solid lines): (a) Herdan's C, C_H, and (b+c) z-score, Eq. (4.28), in which we calculate the expected null model, $\mu(M)$, according to Eq. (4.26) with parameters $\gamma = 1.77$, $b = 7873$ (compare Tab. 3.1), and $a \to \infty$ (in b) or $a = 0.08$ (in c). The solid lines are contours corresponding to values of $V(N)$ that yield the same measure of vocabulary richness varying from rich (red: $C_H = 0.98$ and $z = 4$) to poor (purple: $C_H = 0.8$ and $z = -4$) vocabulary. The article with the richest vocabulary according to each measure is marked by $\times$ (red).

ones described above are needed. In Fig. 4.4 we compare the accuracy of measures of vocabulary richness according to Herdan's C, C_H, Fig. 4.4(a), and the z-score, Fig. 4.4(b+c). For the latter, we use Eq. (4.28) and calculate $\mu(N)$ from Poisson word usage by fixing Zipf's law and assuming Gamma-distributed word frequencies across documents, see Eq. (4.26). We see in Fig. 4.4(a) that Herdan's C, C_H, shows a strong bias towards assigning high values of C_H to shorter texts: following a line with constant C_H we observe for $N \gtrsim 10$ articles with a vocabulary below average while for $N > 1000$ articles with a vocabulary above average. A similar (weaker) bias is observed in Fig. 4.4(b) for the calculation of the z-score for the case in which we consider deviations from the pure Heaps' law but treat frequencies of individual words as fixed, i.e. ignoring topicality. The z-score calculations including topicality in Fig. 4.4(c) show that we obtain a measure of vocabulary richness which is approximately unbiased with respect to the text length N (contour lines are roughly horizontal). Furthermore, in contrast to the two other measures, we correctly assign the highest z-score to the article with the highest ratio $V(N)/V_{\rm avg}(N)$. Altogether, this implies that it is not only important to take into account deviations from the pure Heaps' law but that it is crucial to consider topicality in the form of a quenched average.

4.3. Comparing word frequency distributions

In this section we quantify the distance between two observed instances of text based on the whole ensemble of words and their frequency of usage using tools from information theory.

As a motivational example, in Fig. 4.5 we show the word-frequency distribution of the Google-ngram database from 1850, 1900, and 1950. We see that the distribution itself remains essentially the same, a fat-tailed Zipf distribution

$$p(r) \propto r^{-\gamma}, \tag{4.29}$$

where p is the frequency of the r-th most frequent word and $\gamma \gtrapprox 1$ (we refer to Sec. 3.1 for a detailed analysis of the rank-frequency distribution in the Google-ngram database). However, changes are seen in the frequency p (or rank) of specific words, e.g., *ship* lost and *genetic* won popularity. Measures that quantify such changes are essential to answer questions such as: Is the vocabulary from 1900 more similar to the one from 1850 or to the one from 1950? How similar are two vocabularies (e.g., from different years)? Are the two finite-size observations compatible with a finite sample of the same underlying vocabulary? How similar are the vocabulary of different authors or disciplines? How fast is the lexical change taking place?

Heavy-tailed and broad distributions of symbol frequencies are not only typical in natural languages but appear also in the DNA (n-grams of base pairs for large n) [MBG+94], in gene expression [FK03], and music [BCN08, SCBn+12]. The slow decay observed in a broad range of frequencies implies that there is no typical frequency for words and therefore relevant changes can occur in different ranges

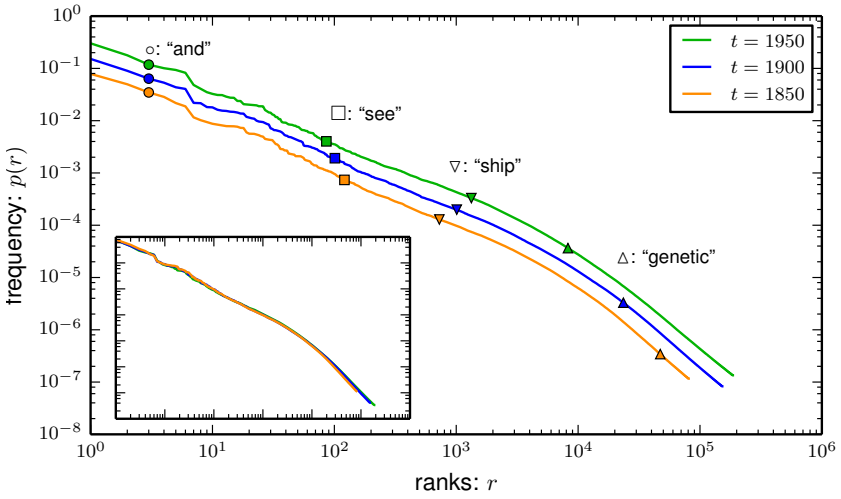

Figure 4.5.: The English vocabulary in different years. Rank-frequency distribution $p(r)$ of individual years t for $t = 1850$, 1900, and 1950 of the Google-ngram database, multiplied by a factor of 1, 2, and 4, respectively, for better visual comparison. The inset shows the original un-transformed data (same axis), highlighting that the rank-frequency distributions are almost indistinguishable. Individual words (e.g. "and","see","ship","genetic") show changes in rank and frequency (symbols), where words with larger ranks (i.e. smaller frequencies) show larger change.

of the p-spectrum, from the few large-frequency words all the way to the many low-frequency words. This imposes a challenge to define similarity measures that are able to account for this variability and that also yield accurate estimations based on finite-size observations.

In Sec. 4.3.1 we quantify the vocabulary similarity using a spectrum of measures D_α based on the generalized entropy of order α ($D_{\alpha=1}$ recovers the usual Jensen-Shannon divergence). In Sec. 4.3.2 we show how varying α magnifies differences in the vocabulary at different scales of the (fat-tailed) frequency spectrum, thus providing different information on the vocabulary change. In Sec. 4.3.3 we show the problem of measuring these divergences in samples of finite size. In particular, for the case of fat-tailed distributions, the large number of low-frequency symbols hinder an accurate finite-size estimation of the corresponding entropies.

4.3.1. Definition

Consider the probability distribution $\boldsymbol{p} = (p_1, p_2, \ldots, p_S)$ of a random variable over a discrete, countable set of symbols $i = 1, \ldots, S$ (where later we include the possibility for $S \to \infty$). One theoretically sound and natural measure coming from information theory for quantifying the difference between two such probability distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ is the Jensen-Shannon divergence (JSD) [Lin91]

$$D(\boldsymbol{p}, \boldsymbol{q}) = H\left(\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) - \frac{1}{2}H(\boldsymbol{p}) - \frac{1}{2}H(\boldsymbol{q}), \tag{4.30}$$

where H is the Shannon entropy [CT06]

$$H(\boldsymbol{p}) = -\sum_i p_i \log p_i. \tag{4.31}$$

This definition has several properties which are useful in the interpretation as a distance: i) $D(\boldsymbol{p}, \boldsymbol{q}) \geq 0$ where the equality holds if and only if $\boldsymbol{p} = \boldsymbol{q}$; ii) $D(\boldsymbol{p}, \boldsymbol{q}) = D(\boldsymbol{q}, \boldsymbol{p})$ (it is a symmetrized Kullback-Leiber divergence [Lin91]); iii) $\sqrt{D(\boldsymbol{p}, \boldsymbol{q})}$ fulfills the triangle inequality and thus is a metric [ES03]; and iv) $D(\boldsymbol{p}, \boldsymbol{q})$ equals the mutual information of variables sampled from $\boldsymbol{p}$ and $\boldsymbol{q}$, i.e., $D(\boldsymbol{p}, \boldsymbol{q})$ equals the average amount of information in one randomly sampled word-token about which of the two distribution it was sampled from [GBGC+02]. The JSD is widely used in the statistical analysis of language [MS99], e.g. to automatically find individual documents that are (semantically) related [BND+11, MKEHG11] or to track the rate of evolution in the lexical inventory of a language over historical time scales [BSW14, PDD15b].

Here we also consider the generalization of JSD in which H in Eq. (4.31) is replaced by the generalized entropy of order α [HC67]

$$H_\alpha(\boldsymbol{p}) = \frac{1}{1-\alpha}\left(\sum_i p_i^\alpha - 1\right), \tag{4.32}$$

yielding a spectrum of divergence measures D_α parameterized by α, first introduced in Ref. [BR82].

The usual JSD is retrieved for $\alpha = 1$. The suitability of Eq. (4.32) to describe physical systems is the subject of investigation of non-extensive statistical mechanics as first proposed in Ref. [Tsa88]. While similar generalizations can be achieved with other formulations of generalized entropies such as the Renyi entropy [Rén61], the corresponding divergence can become negative. In contrast, D_α is strictly non-negative and it was recently shown that $\sqrt{D_\alpha(\boldsymbol{p}, \boldsymbol{q})}$ is a metric for any $\alpha \in (0, 2]$ [BH09]. For heavy-tailed distributions, Eq. (4.29), $H_\alpha < \infty$ for $\alpha > 1/\gamma$.

We define a normalized version of D_α as

$$\tilde{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) = \frac{D_\alpha(\boldsymbol{p}, \boldsymbol{q})}{D_\alpha^{\max}(\boldsymbol{p}, \boldsymbol{q})} \tag{4.33}$$

where

$$D_\alpha^{\max}(\boldsymbol{p}, \boldsymbol{q}) = \frac{2^{1-\alpha} - 1}{2} \left(H_\alpha(\boldsymbol{p}) + H_\alpha(\boldsymbol{q}) + \frac{2}{1-\alpha} \right). \tag{4.34}$$

is the maximum possible D_α between $\boldsymbol{p}$ and $\boldsymbol{q}$ obtained assuming that the the set of symbols in each distribution (i.e., the support of $\boldsymbol{p}$ and $\boldsymbol{q}$) are disjoint. The main motivation for using the measure (4.33) is that $\tilde{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) \in [0, 1]$, while the range of admissible values of D_α depends on α. This allows for a meaningful comparison of the divergences $\tilde{D}_\alpha(\boldsymbol{p}, \boldsymbol{q})$ and $\tilde{D}_{\alpha'}(\boldsymbol{p}, \boldsymbol{q})$ for $\alpha \neq \alpha'$ and therefore also for the full spectrum of α's. In general, the metric properties of D_α are not preserved by $\tilde{D}_\alpha$. An exception is the case in which the frequency distribution $p(r)$ underlying all $\boldsymbol{p}$'s and $\boldsymbol{q}$'s is invariant (see Fig. 4.5). Noting that Eq. (4.34) is independent of the symbols we obtain that $D_\alpha^{\max}(\boldsymbol{p}, \boldsymbol{q})$ is a constant for all $\boldsymbol{p}$'s and $\boldsymbol{q}$'s and therefore the metric property is preserved for $\tilde{D}_\alpha$.

4.3.2. Interpretation

In order to clarify the interpretation of D_α, it is useful to consider a toy model. As in Fig. 4.5, we consider two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ that have exactly the same frequency distribution $p(r)$ but differ in (a subset of) the symbols they use. For simplicity, we consider that symbols that differ in the two cases appear only in one of the distributions. More precisely, denoting by $I_p = \{A, B, C, D, E, \ldots\}$ the set of symbols in $\boldsymbol{p}$ with probabilities p_i and $i \in I_p$, we replace a subset $I^* \subset I_p$ of symbols in $\boldsymbol{q}$ by a new symbol with the same probability (this ensures that the frequency distribution is conserved). Thus the set of symbols in $\boldsymbol{q}$ is $I_q = \{i | i \in I_p \setminus I^*\} \cup \{i^\dagger | i \in I^*\}$ with $p_i = q_i$ for $i \in I_p \setminus I^*$ and $p_i = q_{i^\dagger}$ for $i \in I^*$, see Fig. 4.6 for one example.

For a given distribution $\boldsymbol{p}$ and a set of replaced symbols I^*, we compute $D_\alpha(\boldsymbol{p}, I^*) \equiv D_\alpha(\boldsymbol{p}, \boldsymbol{q})$ as

$$D_\alpha(\boldsymbol{p}, I^*) = c_\alpha \sum_{i \in I^*} p_i^\alpha, \tag{4.35}$$

where $c_\alpha = (2^{(1-\alpha)} - 1)/(1 - \alpha)$. The maximum is given by

$$D_\alpha^{\max}(\boldsymbol{p}, I^*) = c_\alpha \sum_{i \in I_p} p_i^\alpha \tag{4.36}$$

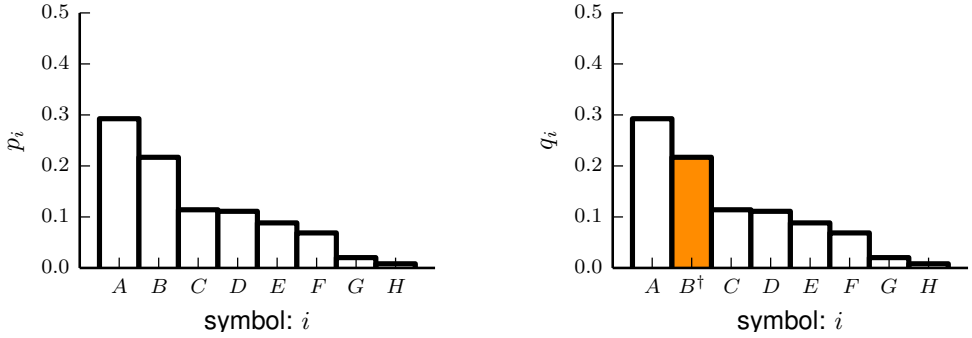

Figure 4.6.: Illustration of our toy model where $\boldsymbol{p}$ (left) and $\boldsymbol{q}$ (right) have the same rank-frequency distribution, but differ in the probability for individual symbols. In this example, $\boldsymbol{p}$ and $\boldsymbol{q}$ are the same ($p_i = q_i$) for $i \in \{A, C, D, E, F, G, H\}$, while the symbol $i = B$ in $\boldsymbol{p}$ is replaced by $i = B^\dagger$ in $\boldsymbol{q}$ with $p_{i=B} = q_{i=B^\dagger}$ and $p_{i=B^\dagger} = q_{i=B} = 0$.

such that

$$\tilde{D}_\alpha(\boldsymbol{p}, I^*) = \frac{\sum_{i \in I^*} p_i^\alpha}{\sum_{i \in I_p} p_i^\alpha}. \tag{4.37}$$

This shows that each symbol $i \in I^*$ that is replaced by a new symbol contributes p_i^α to D_α. It is thus clear that varying α, the contribution of different frequencies become magnified (e.g. for $\alpha \gg 1$ large frequencies are enhanced while for $\alpha < 0$ low frequencies contribute more to D_α than large frequencies).

In particular, for $\alpha = 0$, $\tilde{D}_{\alpha=0}(\boldsymbol{p}, I^*) = \frac{|I^*|}{|I_p|}$ is the fraction of symbols (types) that are different in $\boldsymbol{p}$ and $\boldsymbol{q}$. Each symbol i counts the same irrespective of their probabilities p_i. For $\frac{|I^*|}{|I_p|} \ll 1$, $\tilde{D}_{\alpha=0}(\boldsymbol{p}, I^*) = 1 - J(I_p, I_q)$, where $J(I_p, I_q) = \frac{|I_p \cap I_q|}{|I_p \cup I_q|}$ is the Jaccard-coefficient between the two sets I_p and I_q, an ad-hoc defined similarity measure widely used in information retrieval [MS99]. For $\alpha = 1$, $\tilde{D}_{\alpha=1}(\boldsymbol{p}, I^*) = \sum_{i \in I^*} p_i$ showing that each replaced symbol is weighted by its probability p_i and thus that $\tilde{D}_{\alpha=1}$ measures the distance in terms of tokens.

The full spectrum $\tilde{D}_\alpha$ offers information on changes in all frequencies, a point which is particularly important for the case of fat-tailed distributions because word frequencies vary over many orders of magnitude. Figure 4.7 illustrates how different values of α are able to capture changes at different regions in the frequency spectrum. In particular, it shows that $\tilde{D}_\alpha$ grows (decays) with α when the modified symbols have high (low) frequency. Furthermore, the comparison between two given changes allow us to conclude about which change was more significant at different regions of the word-frequency spectrum. In the example of the figure, both changes (the two lines) are equally significant from the point of view of the modified tokens ($\tilde{D}_1$ are the same), the change in the left affects more types ($\tilde{D}_0$ is larger), and the change in the right affects more frequent words ($\tilde{D}_\alpha$ is larger for $\alpha \gg 1$).

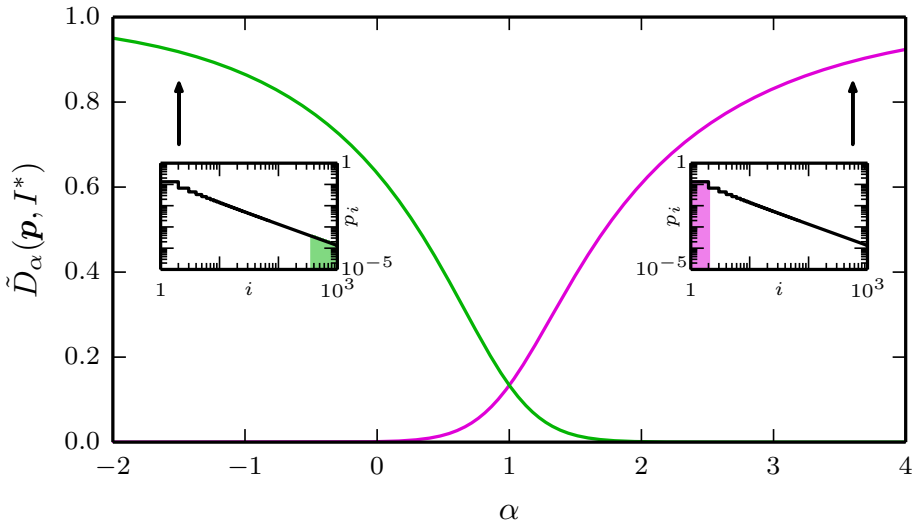

Figure 4.7.: The spectrum $\tilde{D}_\alpha(\boldsymbol{p}, I^*)$ for two different changes. The lines correspond to Eq. (4.37) with $p_i \propto i^{-1}$ with $i = 1, 2, \ldots, 1000$ and two different sets of replaced symbols I_1^*, I_2^*. Right inset: $I_1^* = \{1\}$, i.e., only the symbol with the highest probability, $p_{i=1} \approx 0.13$ is changed. Left inset: $I_2^* = \{368, \ldots, 1000\}$, i.e, the symbols with small probability are replaced. The choice of I_2^* was made such that $\sum_{i \in I_2^*} p_i \approx p_{i=1}$ and therefore $\tilde{D}_{\alpha=1}(\boldsymbol{p}, I_1^*) \approx \tilde{D}_{\alpha=1}(\boldsymbol{p}, I_2^*)$.

4.3.3. Finite-size estimation: Analytical calculations

In this section we turn to the estimation of $\tilde{D}_\alpha$ from data. Even if $\tilde{D}_\alpha$ is defined with respect to distributions $\boldsymbol{p}$ and $\boldsymbol{q}$, Eq. (4.33), in practice these distributions are estimated from sequences with finite size N (total number of symbols or word-tokens) yielding finite-size estimates of the distributions $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$. The main obstacle in obtaining accurate estimates of $\tilde{D}_\alpha$ is that it requires the estimation of entropies for which, in general, unbiased estimators do not exist [Sch04]. Accordingly, even if $\boldsymbol{p} = \boldsymbol{q}$, in practice $H_\alpha(\hat{\boldsymbol{p}}) \neq H_\alpha(\hat{\boldsymbol{q}})$ and $\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) > 0$ are measured not only in single realizations, but also on average (the bias). Besides the bias, we are also interested in the expected fluctuation (standard deviation) of the estimations of H_α and $\tilde{D}_\alpha$ and how both they depend on the sequence size N for large N. In fat-tailed distributions such as Eq. (4.29), these estimations are based on an observed vocabulary V (number of different symbols) that grows sub-linearly with N as [Her60, Hea78] (we refer to Sec. 3.2 for a detailed discussion on the vocabulary growth, i.e. Heaps' law)

$$V(N) \propto N^{1/\gamma}. \tag{4.38}$$

This implies that the entropies in Eq. (4.32) are estimated based on a sum of $V \to \infty$ terms (for $N \to \infty$). In practice, γ and the precise functional form of the fat-tailed distribution are unknown and therefore, besides $\tilde{D}_\alpha$, the estimation of H_α is also of interest (see Ref. [dW99] for the case in which a power law form of $\boldsymbol{p}$ is assumed to be known a priori).

Here we extend and generalize previous results [Mil55, Bas59, Har75, HSE94] to arbitrary α. Given a probability distribution $\boldsymbol{p}$ and the measured probabilities $\hat{\boldsymbol{p}}$ from a finite sample of N word-tokens,

we expand $H_\alpha(\hat{\boldsymbol{p}})$ around the true probabilities p_i up to second order as

$$H_\alpha(\hat{\boldsymbol{p}}) \approx H_\alpha(\boldsymbol{p}) + \sum_{i:\hat{p}_i>0} (\hat{p}_i - p_i)\frac{\alpha}{1-\alpha} p_i^{\alpha-1} - \frac{1}{2}\sum_{i:\hat{p}_i>0} (\hat{p}_i - p_i)^2 \alpha p_i^{\alpha-2} \tag{4.39}$$

where we used that $\frac{\partial H_\alpha}{\partial p_i} = \alpha/(1-\alpha) p_i^{\alpha-1}$ and $\frac{\partial^2 H}{\partial p_i \partial p_j} = -\alpha p_i^{\alpha-2}\delta_{i,j}$. We then calculate $\mathbb{E}\left[H_\alpha(\hat{\boldsymbol{p}})\right]$ by averaging over the different realization of the random variables $\hat{p}_i$ by assuming that the absolute frequency of each symbol i is drawn from an independent binomial with probability p_i such that $\mathbb{E}\left[\hat{p}_i\right] = p_i$ and $\mathbb{V}\left[\hat{p}_i\right] = p_i(1-p_i)/N \approx p_i/N$ yielding

$$\mathbb{E}\left[H_\alpha(\hat{\boldsymbol{p}})\right] \approx H_\alpha(\boldsymbol{p}) - \frac{\alpha}{2N}\sum_{i\in V} p_i^{\alpha-1} = H_\alpha(\boldsymbol{p}) - \frac{\alpha V^{(\alpha)}}{2N}, \tag{4.40}$$

which defines the vocabulary size of order α

$$V^{(\alpha)} \equiv \sum_{i\in V} p_i^{\alpha-1}. \tag{4.41}$$

From Eq. (4.40) we see that the bias in the entropy estimation $|H_\alpha(\boldsymbol{p}) - \mathbb{E}\left[H_\alpha(\hat{\boldsymbol{p}})\right]|$ depends only on $V^{(\alpha)}$ and N. Similar calculations (see the end of this section) show that the large N behavior of the bias and the fluctuations (variance) of H_α, D_α, and $\tilde{D}_\alpha$ can be written as simple functions of $V^{(\alpha)}$ and N, as summarized in Tab. 4.1.

	H_α	$D_\alpha, \tilde{D}_\alpha(\boldsymbol{p}\neq\boldsymbol{q})$	$D_\alpha, \tilde{D}_\alpha(\boldsymbol{p}=\boldsymbol{q})$
Bias:	$V^{(\alpha)}/N$	$V^{(\alpha)}/N$	$V^{(\alpha)}/N$
Fluctuations:	$V^{(2\alpha)}/N$	$V^{(2\alpha)}/N$	$V^{(2\alpha-1)}/N^2$

Table 4.1.: Scaling of the bias $|\mathbb{E}[\hat{X}] - X|$ and the fluctuations $\mathbb{V}[X] \equiv \mathbb{E}[\hat{X}^2] - \mathbb{E}[\hat{X}]^2$ of estimations $\hat{X}$. The results are valid for large sequence sizes N and depend on the vocabulary of order α, $V^{(\alpha)}$ as in Eqs. (4.41) and (4.42). Results are shown for $X = H_\alpha$ [order α entropy, Eq. (4.32)], D_α [generalized divergence], $\tilde{D}_\alpha$ [normalized divergence, Eq. (4.33)], see end of this section for the derivations. For $\tilde{D}_\alpha$, we approximate $\tilde{D}_\alpha \approx D_\alpha/\mathbb{E}[D_\alpha^{\max}]$.

We now focus on the dependence of $V^{(\alpha)}$ on N. The sum $\sum_{i\in V}$ in Eq. (4.41) indicates that in N samples, on average, $V = V(N) \equiv V^{(\alpha=1)}$ different symbols are observed. If for $N \to \infty$ the vocabulary V converges to a finite value, $V^{(\alpha)}$ in Eq. (4.41) also converges and the bias scales as $1/N$. A more interesting scenario happens when V grows with N. For the fat-tailed case of interest here, V grows as $N^{1/\gamma}$, Eq. (4.38), and we obtain (see end of this section) that $V^{(\alpha)}$ scales for large N as

$$V^{(\alpha)} \propto \begin{cases} N^{-\alpha+1+1/\gamma}, & \alpha < 1+1/\gamma, \\ \text{constant}\ , & \alpha > 1+1/\gamma, \end{cases} \tag{4.42}$$

where $\gamma > 1$ is the Zipf exponent defined in Eq. (4.29) and α is the order of the entropy in Eq. (4.32).

From the combination of Eq. (4.42) and Tab. 4.1 we obtain the scalings with sequence size N of

the estimators of H_α, D_α, and $\tilde{D}_\alpha$ in a fat-tailed distribution with exponent γ. These scalings are summarized in Tab. 4.2. Three scaling regimes can be identified for the bias and for the fluctuations. (i) For large α, the decay is $1/N$ (except when $\boldsymbol{p} = \boldsymbol{q}$, where the fluctuations decay even faster as $1/N^2$) as in the case of a finite vocabulary and short-tailed distributions. (ii) For intermediate α, a sub-linear decay with N is observed. This regime appears exclusively in fat-tailed distributions and has important consequences in real applications, as shown below. From the exponents of the sub-linear decay we see that the bias decays more slowly than the fluctuations. (iii) For small α, $\alpha < 1/\gamma$, $H_\alpha(\boldsymbol{p})$ is not defined thus the estimator for the mean of H_α and D_α diverge. The growth of H_α (and therefore $D_\alpha^{\max}$) and D_α with N have the same scaling and therefore cancel each other for $\tilde{D}_\alpha$, in which case a convergence to a well defined value is found (the fluctuation of $\tilde{D}_\alpha$ still decays in this regime).

Finite size estimation of H_α, D_α, and $\tilde{D}_\alpha$ and derivation of Eq. (4.42)

In the following paragraphs we present the full calculations for the mean (i.e. the bias) and the fluctuations in finite-size estimates of H_α, D_α, and $\tilde{D}_\alpha$ as well as the derivation of Eq. (4.42). The starting point is a finite sample $\hat{\boldsymbol{p}} = (n_1/N, n_2/N, \ldots, n_V/N)$ of size N (where n_i is the number of times symbol i was observed) which we assume is obtained from N identical and independent draws

	$\mathbb{E}[H_\alpha(\hat{\boldsymbol{p}})]$	$\mathbb{E}[D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})]$	$\mathbb{E}[\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})]$
$\alpha_1^{\mathbb{E}}$	$1/\gamma$	$1/\gamma$	$1/\gamma$
$\alpha_2^{\mathbb{E}}$	$1 + 1/\gamma$	$1 + 1/\gamma$	$1 + 1/\gamma$
$\alpha < \alpha_1^{\mathbb{E}}$	$cN^{-\alpha+1/\gamma}$	$cN^{-\alpha+1/\gamma}$	c
$\alpha_1^{\mathbb{E}} < \alpha < \alpha_2^{\mathbb{E}}$	$H_\alpha(\boldsymbol{p}) + cN^{-\alpha+1/\gamma}$	$D_\alpha(\boldsymbol{p}, \boldsymbol{q}) + cN^{-\alpha+1/\gamma}$	$\tilde{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) + cN^{-\alpha+1/\gamma}$
$\alpha > \alpha_2^{\mathbb{E}}$	$H_\alpha(\boldsymbol{p}) + cN^{-1}$	$D_\alpha(\boldsymbol{p}, \boldsymbol{q}) + cN^{-1}$	$\tilde{D}_\alpha(\boldsymbol{p}, \boldsymbol{q}) + cN^{-1}$

	$\mathbb{V}[H_\alpha(\hat{\boldsymbol{p}})]$	$\mathbb{V}[D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})]$		$\mathbb{V}[\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})]$	
		$\boldsymbol{p} \neq \boldsymbol{q}$	$\boldsymbol{p} = \boldsymbol{q}$	$\boldsymbol{p} \neq \boldsymbol{q}$	$\boldsymbol{p} = \boldsymbol{q}$
$\alpha_1^{\mathbb{V}}$	$1/(2\gamma)$	$1/(2\gamma)$	$1/(2\gamma)$	$1/\gamma$	$1/\gamma$
$\alpha_2^{\mathbb{V}}$	$\frac{1}{2}(1 + 1/\gamma)$	$\frac{1}{2}(1 + 1/\gamma)$	$1 + 1/(2\gamma)$	$\frac{1}{2}(1 + 1/\gamma)$	$1 + 1/(2\gamma)$
$\alpha < \alpha_1^{\mathbb{V}}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-1/\gamma}$	$cN^{-1/\gamma}$
$\alpha_1^{\mathbb{V}} < \alpha < \alpha_2^{\mathbb{V}}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-2\alpha+1/\gamma}$	$cN^{-2\alpha+1/\gamma}$
$\alpha > \alpha_2^{\mathbb{V}}$	cN^{-1}	cN^{-1}	cN^{-2}	cN^{-1}	cN^{-2}

Table 4.2.: Summary of finite size scaling for distributions with fat tails. Mean ($\mathbb{E}$) and variance ($\mathbb{V}$) of the plug-in estimator of H_α, D_α, and $\tilde{D}_\alpha$ for samples $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$ each of size N drawn randomly from $\boldsymbol{p}$ and $\boldsymbol{q}$ with power law rank-frequency distributions with exponent $\gamma > 1$, Eq. (4.29). The results are obtained combining Tab. 4.1 with Eq. (4.42) (for details see end of this section). The constant c depends on α and has a different value in each case but is independent of N. The limit $\gamma \to \infty$ corresponds to the case in which both $\boldsymbol{p}$ and $\boldsymbol{q}$ have short tails.

from the distribution $\boldsymbol{p}$ giving an estimator for H_α:

$$H_\alpha(\hat{\boldsymbol{p}}) = \frac{1}{1-\alpha}\left(\sum_{i:\hat{p}_i>0} \hat{p}_i^\alpha - 1\right). \tag{4.43}$$

In order to take the corresponding expectation values we expand $\hat{p}_i^\alpha$ around the true probabilities p_i up to second order

$$\hat{p}_i^\alpha \approx p_i^\alpha + (\hat{p}_i - p_i)\alpha p_i^{\alpha-1} + \frac{1}{2}(\hat{p}_i - p_i)^2\alpha(\alpha-1)p_i^{\alpha-2} \tag{4.44}$$

and average over the realizations of the random variables $\hat{p}_i^\alpha$ by assuming that each symbol is drawn independently from binomial with probability p_i such that $\langle(\hat{p}_i - p_i)\rangle = 0$ and $\langle(\hat{p}_i - p_i)^2\rangle = p_i(1-p_i)/N \approx p_i/N$ yielding [HSE94]

$$\langle\hat{p}_i^\alpha\rangle \approx p_i^\alpha + \frac{1}{2N}\alpha(\alpha-1)p_i^{\alpha-1}. \tag{4.45}$$

Estimating H_α

Combining Eqs. (4.43) and (4.45) we obtain for the mean

$$\begin{aligned}
\mathbb{E}[H_\alpha(\hat{\boldsymbol{p}})] \equiv \langle H_\alpha(\hat{\boldsymbol{p}})\rangle &= \frac{1}{1-\alpha}\left(\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle}\langle\hat{p}_i^\alpha\rangle - 1\right) \\
&= \frac{1}{1-\alpha}\left(\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle} p_i^\alpha - 1\right) - \frac{\alpha}{2N}\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle} p_i^{\alpha-1} \\
&= \frac{1}{1-\alpha}\left(V_{\hat{\boldsymbol{p}}}^{(\alpha+1)} - 1\right) - \frac{\alpha}{2N}V_{\hat{\boldsymbol{p}}}^{(\alpha)}
\end{aligned} \tag{4.46}$$

where we introduce the notation $\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle}$ indicating that we average only over the expected number of observed symbols $\langle V_{\hat{\boldsymbol{p}}}\rangle$ in samples $\hat{\boldsymbol{p}}$.

For the variance we get

$$\begin{aligned}
\mathbb{V}[H_\alpha(\hat{\boldsymbol{p}})] \equiv& \mathbb{E}[H_\alpha(\hat{\boldsymbol{p}})^2] - \mathbb{E}[H_\alpha(\hat{\boldsymbol{p}})]^2 \\
=& \frac{1}{(1-\alpha)^2}\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle}\sum_{j\in\langle V_{\hat{\boldsymbol{p}}}\rangle}\left(\langle\hat{p}_i^\alpha\hat{p}_j^\alpha\rangle - \langle\hat{p}_i^\alpha\rangle\langle\hat{p}_j^\alpha\rangle\right) \\
=& \frac{\alpha^2}{(1-\alpha)^2 N}\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle} p_i^{2\alpha-1} - \frac{\alpha^2}{4N^2}\sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle} p_i^{2\alpha-2} \\
=& \frac{\alpha^2}{(1-\alpha)^2}\frac{V_{\hat{\boldsymbol{p}}}^{(2\alpha)}}{N} - \frac{\alpha^2}{4}\frac{V_{\hat{\boldsymbol{p}}}^{(2\alpha-1)}}{N^2}
\end{aligned} \tag{4.47}$$

where we used that two different symbols $i \neq j$ are independently drawn, thus $\sum_{i,j}\langle\hat{p}_i^\alpha\hat{p}_j^\alpha\rangle =$

$\sum_{i\neq j}\langle\hat{p}_i^\alpha\rangle\langle\hat{p}_j^\alpha\rangle + \sum_i\langle\hat{p}_i^{2\alpha}\rangle$.

Estimating D_α

For D_α we have two samples $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$ each of size N randomly sampled from the distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ such that we can express the mean and the variance from the expectation values of the corresponding individual entropies.

Introducing the notation $\boldsymbol{P} \equiv \frac{1}{2}(\boldsymbol{p}+\boldsymbol{q})$ we get for the mean

$$\begin{aligned}\mathbb{E}[D_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})] =& \mathbb{E}[H_\alpha(\hat{\boldsymbol{P}})] - \frac{1}{2}\mathbb{E}\left[H_\alpha\left(\hat{\boldsymbol{p}}\right)\right] - \frac{1}{2}\mathbb{E}\left[H_\alpha\left(\hat{\boldsymbol{q}}\right)\right] \\ =& \frac{1}{1-\alpha}\left\{V_{\hat{\boldsymbol{P}}}^{(\alpha+1)} - \frac{1}{2}V_{\hat{\boldsymbol{p}}}^{(\alpha+1)} - \frac{1}{2}V_{\hat{\boldsymbol{q}}}^{(\alpha+1)}\right\} + \frac{\alpha}{2N}\left\{\frac{1}{2}V_{\hat{\boldsymbol{p}}}^{(\alpha)} + \frac{1}{2}V_{\hat{\boldsymbol{q}}}^{(\alpha)} - \frac{1}{2}V_{\hat{\boldsymbol{P}}}^{(\alpha)}\right\}.\end{aligned} \tag{4.48}$$

where $V_{\hat{\boldsymbol{P}}}^{(\alpha)}$ denotes the generalized vocabulary, Eq. (4.41), for the combined sequence $\hat{\boldsymbol{P}} = \frac{1}{2}(\hat{\boldsymbol{p}}+\hat{\boldsymbol{q}})$, which is of length $2N$.

For the variance we get

$$\begin{aligned}\mathbb{V}[D_\alpha\left(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}}\right)] \equiv& \mathbb{E}[D_\alpha\left(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}}\right)^2] - \mathbb{E}[D_\alpha\left(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}}\right)]^2 \\ =& \mathbb{V}[H_\alpha(\hat{\boldsymbol{P}})] + \frac{1}{4}\mathbb{V}\left[H_\alpha(\hat{\boldsymbol{p}})\right] + \frac{1}{4}\mathbb{V}\left[H_\alpha(\hat{\boldsymbol{q}})\right] - \mathrm{Cov}\left[H_\alpha(\hat{\boldsymbol{P}}), H_\alpha(\hat{\boldsymbol{p}}) + H_\alpha(\hat{\boldsymbol{q}})\right],\end{aligned} \tag{4.49}$$

where $\mathrm{Cov}\left[X,Y\right] \equiv \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. We evaluate the covariance-term in two different ways, i.e.

$$\begin{aligned}&(1-\alpha)^2\mathrm{Cov}\left[H_\alpha(\hat{\boldsymbol{P}}), H_\alpha(\hat{\boldsymbol{p}}) + H_\alpha(\hat{\boldsymbol{q}})\right] \\ =& \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \left(\sum_{j:\hat{p}_j>0} \hat{p}_j^\alpha + \sum_{j:\hat{q}_j>0} \hat{q}_j^\alpha \right) \right\rangle - \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \right\rangle \left(\left\langle \sum_{j:\hat{p}_j>0} \hat{p}_j^\alpha \right\rangle + \left\langle \sum_{j:\hat{q}_j>0} \hat{q}_j^\alpha \right\rangle \right) \\ =& \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \sum_{j:\hat{p}_j+\hat{q}_j>0} \left(\hat{p}_j^\alpha + \hat{q}_j^\alpha\right) \right\rangle - \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \right\rangle \left\langle \sum_{j:\hat{p}_j+\hat{q}_j>0} \left(\hat{p}_j^\alpha + \hat{q}_j^\alpha\right) \right\rangle \\ =& \sum_{i\in\langle V_{\hat{\boldsymbol{P}}}\rangle} \left\{ \left\langle \hat{P}_i^\alpha \left(\hat{p}_i^\alpha + \hat{q}_i^\alpha\right) \right\rangle - \left\langle \hat{P}_i^\alpha \right\rangle \left(\langle\hat{p}_i^\alpha\rangle + \langle\hat{q}_i^\alpha\rangle\right) \right\}\end{aligned} \tag{4.50}$$

and

$$\begin{aligned}&(1-\alpha)^2\mathrm{Cov}\left[H_\alpha(\hat{\boldsymbol{P}}), H_\alpha(\hat{\boldsymbol{p}}) + H_\alpha(\hat{\boldsymbol{q}})\right] \\ =& \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \sum_{j:\hat{p}_j>0} \hat{p}_j^\alpha \right\rangle - \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \right\rangle \left\langle \sum_{j:\hat{p}_j>0} \hat{p}_j^\alpha \right\rangle \\ &+ \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \sum_{j:\hat{q}_j>0} \hat{q}_j^\alpha \right\rangle - \left\langle \sum_{i:\hat{p}_i+\hat{q}_i>0} \hat{P}_i^\alpha \right\rangle \left\langle \sum_{j:\hat{q}_j>0} \hat{q}_j^\alpha \right\rangle \\ =& \sum_{i\in\langle V_{\hat{\boldsymbol{p}}}\rangle} \left\{ \left\langle \hat{P}_i^\alpha \hat{p}_i^\alpha \right\rangle - \left\langle \hat{P}_i^\alpha \right\rangle \langle\hat{p}_i^\alpha\rangle \right\} + \sum_{i\in\langle V_{\hat{\boldsymbol{q}}}\rangle} \left\{ \left\langle \hat{P}_i^\alpha \hat{q}_i^\alpha \right\rangle - \left\langle \hat{P}_i^\alpha \right\rangle \langle\hat{q}_i^\alpha\rangle \right\}\end{aligned} \tag{4.51}$$

Similarly as in Eq. (4.45) we can approximate

$$\begin{aligned}\left\langle \hat{P}_i^\alpha \right\rangle \approx & P_i^\alpha + \frac{\alpha(\alpha-1)}{4N} P_i^{\alpha-1}, \\ \left\langle \hat{P}_i^\alpha \hat{p}_i^\alpha \right\rangle \approx & P_i^\alpha p_i^\alpha + \frac{\alpha}{4N}(3\alpha-1) P_i^{\alpha-1} p_i^\alpha + \frac{\alpha}{2N}(\alpha-1) P_i^\alpha p_i^{\alpha-1}, \\ \left\langle \hat{P}_i^\alpha \hat{q}_i^\alpha \right\rangle \approx & P_i^\alpha q_i^\alpha + \frac{\alpha}{4N}(3\alpha-1) P_i^{\alpha-1} q_i^\alpha + \frac{\alpha}{2N}(\alpha-1) P_i^\alpha q_i^{\alpha-1}.\end{aligned} \tag{4.52}$$

From this we get for the variance of D_α

$$\begin{aligned}\mathbb{V}[D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})] = & \sum_{i \in \langle V_{\hat{\boldsymbol{P}}} \rangle} \left\{ \frac{\alpha^2}{(1-\alpha)^2} \frac{1}{2N} P_i^{\alpha-1} \left[P_i^\alpha - \frac{1}{2}\left(p_i^\alpha + q_i^\alpha\right) \right] - \frac{\alpha^2}{16N^2} P_i^{\alpha-1} \left[P_i^{\alpha-1} - \left(p_i^{\alpha-1} + q_i^{\alpha-1}\right) \right] \right\} \\ & + \frac{1}{2} \sum_{i \in \langle V_{\hat{\boldsymbol{p}}} \rangle} \left\{ \frac{\alpha^2}{(1-\alpha)^2} \frac{1}{2N} p_i^\alpha \left[p_i^{\alpha-1} - P_i^{\alpha-1} \right] - \frac{\alpha^2}{8N^2} p_i^{\alpha-1} \left[p_i^{\alpha-1} - P_i^{\alpha-1} \right] \right\} \\ & + \frac{1}{2} \sum_{i \in \langle V_{\hat{\boldsymbol{q}}} \rangle} \left\{ \frac{\alpha^2}{(1-\alpha)^2} \frac{1}{2N} q_i^\alpha \left[q_i^{\alpha-1} - P_i^{\alpha-1} \right] - \frac{\alpha^2}{8N^2} q_i^{\alpha-1} \left[q_i^{\alpha-1} - P_i^{\alpha-1} \right] \right\}.\end{aligned} \tag{4.53}$$

Now we can see that for $\boldsymbol{p} = \boldsymbol{q} = \boldsymbol{P}$ we get

$$\mathbb{V}[D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})]_{\boldsymbol{p}=\boldsymbol{q}} = \sum_{i \in \langle V_{\hat{\boldsymbol{P}}} \rangle} \frac{1}{16N^2} \alpha^2 p_i^{2\alpha-2} = \frac{\alpha^2}{16N^2} V_{\hat{\boldsymbol{P}}}^{(2\alpha-1)}. \tag{4.54}$$

While for arbitrary $\boldsymbol{p}$ and $\boldsymbol{q}$ the variance of the D_α contains the variances of the individual entropies (e.g. $V_{\hat{\boldsymbol{P}}}^{(2\alpha)}/N$) and a covariance term, (only) in the special case $\boldsymbol{p} = \boldsymbol{q}$ all first-order terms $(1/N)$ vanish yielding a qualitatively different behaviour $V_{\hat{\boldsymbol{P}}}^{(2\alpha-1)}/N^2$.

Estimating $\tilde{D}_\alpha$

The finite-size estimation of $\tilde{D}_\alpha$ can be obtained approximately by

$$\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) = \frac{D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})}{D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})_{\max}} \approx \frac{D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})}{\mathbb{E}\left[D_\alpha^{\max}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right]} \tag{4.55}$$

such that

$$\begin{aligned}\mathbb{E}\left[\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right] \approx & \frac{\mathbb{E}\left[D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right]}{\mathbb{E}\left[D_\alpha^{\max}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right]}, \\ \mathbb{V}\left[\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right] \approx & \frac{\mathbb{V}\left[D_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right]}{\mathbb{E}\left[D_\alpha^{\max}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})\right]^2}.\end{aligned} \tag{4.56}$$

The mean of $D_\alpha^{\max}$ is given according to Eq. (4.34) as a linear combination of the individual entropies of $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$

$$\mathbb{E}\left[D_\alpha^{\max}(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})\right] = \frac{2^{1-\alpha}-1}{2}\left(\mathbb{E}\left[H_\alpha\left(\hat{\boldsymbol{p}}\right)\right] + \mathbb{E}\left[H_\alpha\left(\hat{\boldsymbol{q}}\right)\right] + \frac{2}{1-\alpha}\right). \tag{4.57}$$

Derivation of Eq. (4.42)

In this paragraph we derive the scaling of the generalized vocabulary $V^{(\alpha)}$ defined in Eq. (4.41) assuming that $\boldsymbol{p}$ is a power law of the form $p_i \propto i^{-\gamma}$, Eq. (4.29). Instead of looking at the probability of individual symbols i, we consider the distribution of frequencies n, which in this case yields $p(n) \propto n^{-1-1/\gamma}$ [New05]. Consider the sum $\sum_{i\in V} p_i = \frac{1}{N}\sum_{i\in V} n_i = \frac{1}{N} S_V(\gamma)$, where $S_V(\gamma)$ corresponds to the sum of V i.i.d. random variables n_i $(i = 1,\ldots,V)$ drawn from the distribution $p(n)$. It can be shown that [BG90]

$$S_V(\gamma) \propto \begin{cases} V^\gamma, & \gamma > 1, \\ V, & \gamma < 1. \end{cases} \tag{4.58}$$

The case $\gamma = 1$ includes additional logarithmic corrections, but is not of relevance for the discussion, therefore, we leave it for sake of simplicity. In the same way, we can treat $\sum_{i\in V} p_i^\mu = \frac{1}{N^\mu}\sum_{i\in V} n_i^\mu = \frac{1}{N^\mu} S_V(\gamma\mu)$ by noting that $S_V(\gamma\mu)$ can be interpreted as the sum of V i.i.d. random variables n_i $(i = 1,\ldots,V)$, where $n_i \sim \tilde{p}(n)$ with $\tilde{p}(n) \propto n^{-1-1/(\gamma\mu)}$ such that we get

$$S_V(\gamma\mu) \propto \begin{cases} V^{\gamma\mu}, & \mu < 1/\gamma, \\ V, & \mu > 1/\gamma. \end{cases} \tag{4.59}$$

By setting $\mu = \alpha - 1$ in Eq. (4.41) and noting that for $p_i \propto i^{-\gamma}$, Eq. (4.29), the number of different symbols scales as $V \propto N^{1/\gamma}$, Eq. (4.38), we obtain Eq. (4.42).

4.3.4. Finite-size estimation: Numerical simulations

Here we perform numerical estimations of the normalized divergence spectrum $\tilde{D}_\alpha$ that illustrate the regimes derived above, confirm the validity of the approximations used in their derivations, and show that the same scalings are observed for different entropy estimators. We sample twice N symbols (tokens) from the same distribution ($\boldsymbol{p} = \boldsymbol{q}$), and therefore $\tilde{D}_\alpha = 0$ and the expected value $\mathbb{E}[\tilde{D}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})]$ is the bias. (The fact that the bias shows a slower decay with N than the fluctuations makes these two effects distinguishable also in this $\tilde{D}_\alpha = 0$ case because $\mathbb{E}[\tilde{D}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})] \gg \mathbb{V}[\tilde{D}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})]$ for large N).

$\tilde{D}_{\alpha=1}$ (case $\boldsymbol{p} = \boldsymbol{q}$)

We start with the most important prediction of our analytical calculations above: the existence in fat-tailed distributions of a regime for which the bias and fluctuations of $\tilde{D}_\alpha$ decay with N more slowly

than $1/N$. This holds already for $\alpha = 1$, i.e., for the usual Jensen-Shannon divergence, previously shown for the bias of $H_{\alpha=1}$ in Ref. [HSE94]. One potential limitation of our analytical calculations is that they are based on the plug-in estimator obtained from replacing the p_i's in the generalized entropies, Eq. (4.32), by the measured frequencies (i.e. $p_i \mapsto \hat{p}_i = N_i/N$, with N_i being the number of observed word-tokens of type i). To test the generality of our results, in the numerical simulations we use four different estimators of the Shannon entropy (i.e., $\alpha = 1$): i) the *Plug-in* estimator; ii) *Miller*'s estimator [Mil55], which takes into account the approximation obtained from the expansion in Eq. (4.40); iii) *Grassberger*'s estimator [Gra08] based on the assumption that frequencies are Poisson-distributed; and iv) a recently proposed *Bayesian* estimator described in [APP14] which is an extension of the approach by Nemenman et al. [NSB02] to the case where the number of possible symbols is unknown or even countably infinite. The numerical results in Fig. 4.8 show that the different estimators are indeed able to reduce the bias of the estimation, but that the scaling of the bias with N remains the same. In particular, the transition from short-tailed to fat-tailed distribution leads to the predicted transition from N^{-1} (N^{-2}) to the slower $N^{-1+1/\gamma}$ ($N^{-2+1/\gamma}$) decay for the bias (fluctuations) for all estimators. The only exception is in the bias of the Bayesian estimator for the exact Zipf's law (4.29), but since this estimator shows a bad performance for the fluctuation and for the real data we conclude that the slower scaling should be expected in general also for this elaborated estimator. These results confirm the generality of our finding that the bias and fluctuation in $\tilde{D}_{\alpha=1}$ decays more slowly than $1/N$ in fat-tailed distributions.

A problem that appears in different contexts is to test whether two finite-size N sequences, described by their empirical distributions $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$, have a common source (null hypothesis). This involves the computation of a single divergence $\tilde{D}_{\alpha=1}(\boldsymbol{p}, \boldsymbol{q})$, which is then compared to the divergence $\tilde{D}_{\alpha=1}(\boldsymbol{p}', \boldsymbol{p}')$ between two finite-size (random) samplings of a single (properly chosen) distribution $\boldsymbol{p}'$ (e.g., $\boldsymbol{p}' = 0.5\boldsymbol{p} + 0.5\boldsymbol{q}$). The probability of observing $\tilde{D}_{\alpha=1}(\boldsymbol{p}', \boldsymbol{p}') \geq \tilde{D}_{\alpha=1}(\boldsymbol{p}, \boldsymbol{q})$ is then reported as a p-value [GBGC$^+$02]. Besides applications in language, e.g. comparing the distribution of word frequencies, this problem appears in the identification of coding- and non-coding regions in DNA [BGGC$^+$00]. The significance of our results is that for the case of fat-tailed distribution the expected $\tilde{D}_{\alpha=1}(\boldsymbol{p}, \boldsymbol{q})$ of the null model may be much larger than the predicted value based on a $1/N$ decay (as observed in short-tailed distributions). If the slower convergence in N is ignored, e.g., by applying standard tests to fat-tailed distributions, one rejects the null hypothesis (low p-value) even if the data is drawn from the same source because the measured distance will be much larger. The example in Fig. 4.8(c) shows that, even when the size of both sequences is on the order of $N \approx 10^5$, the expected $\tilde{D}_{\alpha=1}$ (JSD) is $\mathbb{E}[\tilde{D}_{\alpha=1}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})] \approx 10^{-1}$. This is two orders of magnitude larger than for the exponential distribution in Fig. 4.8(a), where $\mathbb{E}[\tilde{D}_{\alpha=1}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})] \approx 10^{-3}$.

$\tilde{D}_\alpha$ (case $\boldsymbol{p} = \boldsymbol{q}$)

We now consider the estimation of $\tilde{D}_\alpha$ for $\alpha \neq 1$ in the case of fat-tailed distributions (4.29). The numerical results in Fig. 4.9 confirm the existence of the three scaling regimes discussed after Eq. (4.42) and Tab. 4.2. The panels (b) and (d) show the relative reduction in the bias and fluctuations achieved

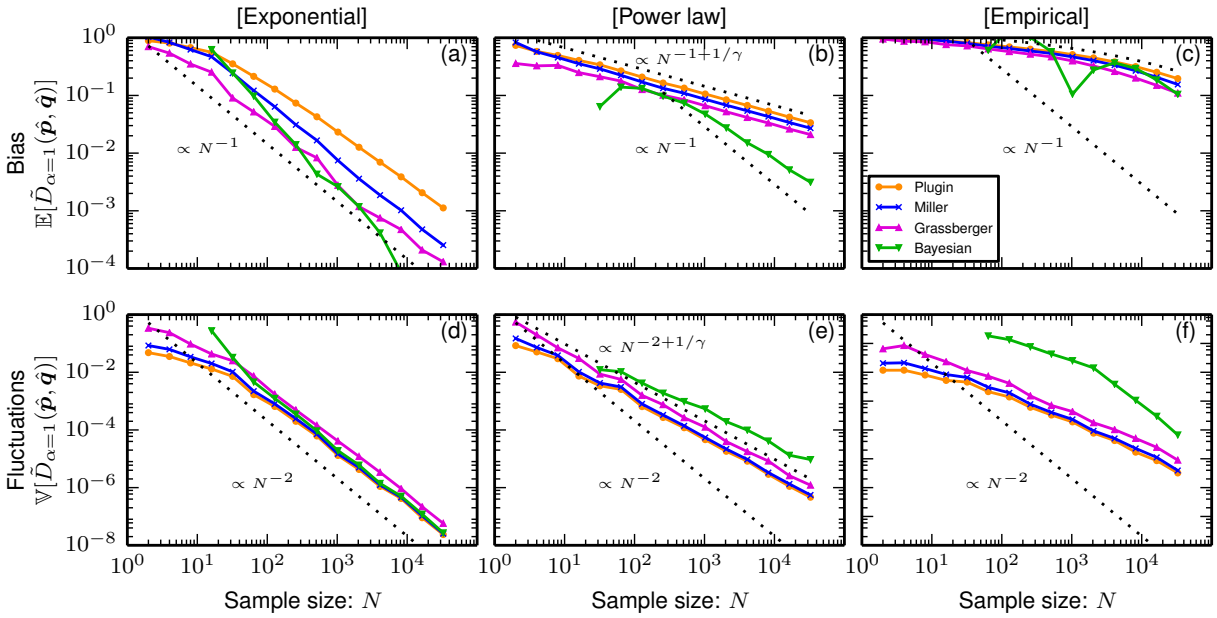

Figure 4.8.: Finite-size estimation of the normalized Jensen-Shannon divergence $\tilde{D} = \tilde{D}_{\alpha=1}$. (a-c) Estimation of $\mathbb{E}[\tilde{D}(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})]$ between two sequences of size N drawn from the same distribution (i.e. $D(\boldsymbol{p},\boldsymbol{q}) = 0$) using four different estimators of the entropy (see text) for three representative distributions: (a) Exponential (short-tailed) distribution $p_i \propto i^{-ai}$ for $i = 0, 1, \ldots$ with $a = 0.1$; (b) Power law (fat-tailed) distribution $p_i \propto i^{-\gamma}$ for $i = 1, 2, \ldots$ with $\gamma = 3/2$; (c) Empirical Zipf-distribution of word frequencies, i.e. rank-frequency distribution $p(r)$ from the complete Google-ngram data, $p_i = f(i = r)$ for $i = 1, \ldots, 4623568$, which is well described by a double power law [GA13]. (d-f) Show the same as (a-c) for the fluctuations $\mathbb{V}[\tilde{D}(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})]$. The dotted lines show the expected scalings from Tab. 4.2 for short-tailed distributions, i.e. N^{-1} (N^{-2}), and power-law distributions, i.e. $N^{-1+1/\gamma}$ ($N^{-2+1/\gamma}$), for the bias (fluctuations). In (c) we show the expected scaling for the bias, $V_{\text{emp}}(N)/N$, where $V_{\text{emp}}(N)$ is the expected number of different symbols in a random sample of size N from the empirical distribution [GA14]. Averages are taken over 100 realizations.

when the sequence size is doubled. For many different α's the relative reduction is larger than 0.5 (0.25) for the bias (fluctuations), a consequence of the slow decay with N that shows the difficulty in obtaining a good estimation of $\tilde{D}_\alpha$. In practice, the exponent γ of the distribution is unknown such that the critical values of α that separate these regimes (e.g. $\alpha_1^{\mathbb{E}} = 1/\gamma$ and $\alpha_2^{\mathbb{E}} = 1 + 1/\gamma$ for the bias) can not be determined a priori. Yet, since $\gamma > 1$, we know that: (i) $\alpha_1^{\mathbb{E}}, \alpha_1^{\mathbb{V}} < 1$ and therefore D_α for $\alpha \geq 1$ is such that $D_\alpha(\boldsymbol{p},\boldsymbol{p}) = 0$ for $N \to \infty$; and (ii) $\alpha_2^{\mathbb{E}}, \alpha_2^{\mathbb{V}} < 2$ and therefore the bias and fluctuations of D_α for $\alpha \geq 2$ decay as $1/N$ (or $1/N^2$ for the fluctuations in the case of $\boldsymbol{p} = \boldsymbol{q}$). This suggests $D_{\alpha=2}$ as a pragmatic choice for empirical measurements because any further increase in α will not lead to a faster convergence.

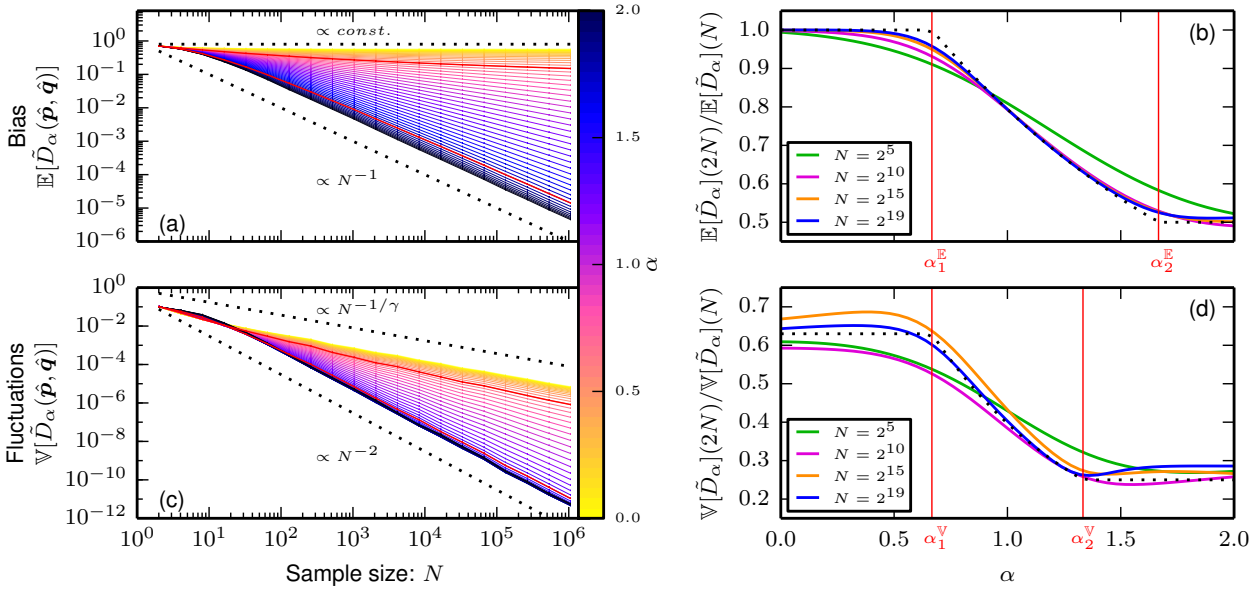

Figure 4.9.: Bias (a,b) and fluctuations (c,d) in finite-size estimation of $\tilde{D}_\alpha$. Estimation of $\mathbb{E}[\tilde{D}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})]$ between two sequences each of size N drawn numerically from the same power law distribution $p_i \propto i^{-\gamma}$ for $i = 1, 2, \ldots, V \to \infty$ with $\gamma = 3/2$ using the plug-in estimator ($p_i \mapsto \hat{p}_i$) for the entropies of order α. (a) Scaling of the bias with N for different α. (b) Decrease of the bias in $\tilde{D}_\alpha$ when sample size is doubled ($N \mapsto 2N$) for different values of N as a function of α. (c) and (d) show the same as (a) and (b) for the fluctuations $\mathbb{V}[\tilde{D}_\alpha(\hat{\boldsymbol{p}},\hat{\boldsymbol{q}})]$. Red lines in all plots indicate the borders between the regimes, $\alpha_1^{\mathbb{E}} = 1/\gamma = 2/3$, $\alpha_2^{\mathbb{E}} = 1 + 1/\gamma = 5/3$ (for the bias in a,b), and $\alpha_1^{\mathbb{V}} = 1/\gamma = 2/3$, $\alpha_2^{\mathbb{V}} = 1 + 1/(2\gamma) = 4/3$ (for the fluctuations in c,d). Dotted lines indicate the predictions based on Tab. 4.1 for $\alpha < \alpha_1^{\mathbb{E}}, \alpha_1^{\mathbb{V}}$ and $\alpha > \alpha_2^{\mathbb{E}}, \alpha_2^{\mathbb{V}}$ (in a,c) and all values of α (in b,d). Averages are taken over 1000 realizations.

5. Modeling topicality

In contrast to the previous chapter, where we showed the necessity to consider topical aspects in order to describe, e.g. the fluctuations in the vocabulary growth, and explored different approaches in quantifying topicality, in this chapter we rather focus on modeling topical aspects in a collection of documents. That is one tries to capture the mesoscopic, i.e. large-scale, structure of the texts in order to obtain a coarse-grained description by identifying coherent groups of documents (and/or words) summarized under the name *topic models*. A main motivation for this approach in the context of information retrieval [MRS08] is to find semantically related documents, i.e. which are similar in content. More generally, it addresses the question of how to organize knowledge accumulated in the form of so-called unstructured information [SB04], e.g. texts. Besides applications in search engines, these approaches are employed in the scientific process itself. Due to the unprecedented growth in the number of scientific papers published, automatically connecting pieces of scientific knowledge has become crucial, especially in biomedical research [RSOH12].

Here, we approach the problem of identifying the large-scale structure of texts by mapping it to the problem of *community detection* in complex networks in the framework of stochastic block models [GPA15]. This not only leads to a more general formulation of the problem at hand but also i) solves many of the intrinsic limitations of topic modeling; ii) leads to an improved understanding of the inference algorithms; and iii) yields much better inference results in terms of model selection.

In Sec. 5.1 we review the basic principles of topic models and community detection. In Sec. 5.2 we show the similarities of and the differences between the two approaches and how community detection methods can be applied to describe the topicality of texts. In Sec. 5.3 we systematically compare state-of-the-art methods from topic models and community detection in artificial and real texts.

5.1. Theoretical framework

In this section we describe briefly the main ideas of topics models and community detection in networks.

5.1.1. Topic models

The term topic models refers to a collection of techniques to describe the variation of word usage across different texts, $d = 1, \ldots, D$. The documents are typically represented as bag-of-words, in which any word-order (e.g. grammar) within a text is disregarded, thus, only counting how often each word $w = 1, \ldots, V$ appears in each document d, $n_{w,d}$. From this, one defines a matrix A, with

elements $A_{w,d} = n_{w,d}$, such that each document is defined as a vector over all words in the vocabulary. However, since the total vocabulary, V, is much larger than the vocabulary of an individual texts, the matrix A is extremely sparse meaning that only very few of the entries in A actually have non-zero entries. In this case, the comparison of individual documents by their vectors in A (e.g. by measuring the Euclidean distance) can be dominated by spurious effects due to the so-called "curse of dimensionality". This motivates approaches based on dimensionality reduction in terms of a number of K latent (i.e. not observable) variables with $K \ll D, V$. In the application to language data, the latent variables are interpreted in terms of the human intuition of topics (e.g. whether a document is about politics or sports) such that they are assumed to capture aspects on the semantic or conceptual level of the individual documents. From this one obtains a coarse-grained description of the individual matrix A, in which each topic is described by the contributions of the words w, and documents are described by the mixture of topics.

Latent semantic indexing

The idea of latent semantic indexing (LSI) [DDF$^+$90] is based on the singular value decomposition (SVD) of the $V \times D$-matrix $A = \{A_{w,d}\}$. The SVD of A can be written in the form

$$A = \Phi \Sigma \Theta^T \tag{5.1}$$

where the matrices Φ and Θ are orthonormal, Σ is diagonal

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{R=\min\{V,N\}} \end{bmatrix}, \tag{5.2}$$

and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{R=\min\{V,D\}}$ are the singular values of the matrix A. The approach in LSI is to approximate A by a partial SVD using only the K largest singular values in Σ

$$\begin{aligned} A \approx A_{(K)} &= \Phi_{(K)} \Sigma_{(K)} \Theta_{(K)}^T \\ &= [\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_K] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_K \end{bmatrix} \begin{bmatrix} \boldsymbol{\Theta}_1^T \\ \vdots \\ \boldsymbol{\Theta}_K^T \end{bmatrix} \end{aligned} \tag{5.3}$$

where $\boldsymbol{\Phi}_k$ is the k-th column of Φ and $\boldsymbol{\Theta}_k^T$ is the k-th row of Θ^T for $k = 1, \ldots, K$. This approximation corresponds to a dimensionality reduction that is rank-K optimal in the sense that $A_{(k)}$ is the matrix of rank K that is closest to the matrix A in terms of the L_2-norm [BDO95]. The $V \times K$-matrix $\Phi_{(k)}$ ($K \times D$-matrix $\Theta_{(k)}^T$) can be interpreted as the projection of the words (documents) into the reduced K-dimensional space which is defined by the K topics obtained from the principal eigenvectors of AA^T in the partial SVD in Eq. (5.3). In this view, each dimension $k = 1, \ldots, K$ is a linear superposition of words given by the columns of $\Phi_{(k)}$ and each document d is a linear superposition of the dimensions

$k = 1, \ldots, K$ given by the rows of $\Theta^T_{(k)}$.

Although LSI is based on the well-understood method of SVD from linear algebra, the methodological foundations remain unsatisfactory. The optimization in terms of the L_2-norm corresponds to an implicit additive-noise assumption [Hof01], which is hard to justify in the application to count data in the form of the matrix A. In addition, the issue of order selection, e.g. choosing the number of dimensions K in LSI, is based on heuristics. The most important drawback is that there is no obvious interpretation of the principal eigenvectors spanning the K-dimensional LSI latent space.

Probabilistic latent semantic indexing

A natural way to address the problems of LSI is to formulate the problem of finding topical structure in a collection of texts in terms of a probabilistic generative process. In this framework, one proposes a model (including a set of parameters) that generates texts with topical structure. In turn, the problem at hand becomes that of statistical inference, i.e. given some observed data A, what are the best parameters in the proposed model. Here one can employ well-established methods from statistics, e.g. Maximum Likelihood, Bayesian inference, or model selection.

One approach by Hofmann [Hof99], probabilistic latent semantic indexing (PLSI), extends LSI to such a probabilistic framework. In PLSI we assume the following generative process for a collection of texts. For each document $d = 1, \ldots, D$ with length n_d and inside each document for each word-token $i_d = 1, \ldots, n_d$:

- a topic z is chosen with probability $P(z_{i_d} = z \mid d) = \theta_{z,d}$,
- a word-type w is chosen with probability $P(w_{i_d} = w \mid z_{i_d} = z) = \phi_{w,z}$.

Therefore, for a randomly chosen word-token, the probability that it belongs to document d and word-type w is

$$P(w, d) = P(d)P(w \mid d) = P(d) \sum_z P(w \mid z)P(z \mid d) = \frac{n_d}{N} \sum_z \phi_{w,z}\theta_{z,w}. \tag{5.4}$$

From this we can write down the likelihood to observe an empirical corpus $A = \{A_{w,d}\}$

$$P(\{A_{w,d}\} \mid \{\phi_{w,z}\}, \{\theta_{z,d}\}) = \prod_d \prod_w P(w, d)^{A_{w,d}} \tag{5.5}$$

and infer $\phi_{w,z}$ and $\theta_{z,d}$ by maximizing the likelihood. The inferred parameters $\phi_{w,z}$ ($\theta_{z,d}$) can thus be interpreted as a dimensionality reduction of A in terms of a decomposition into normalized conditional probability distributions, i.e. $\sum_w \phi_{w,z} = 1$ ($\sum_z \theta_{z,d} = 1$), for the contribution of a word-type w in a topic z (a topic z in a document d).

The PLSI suffers from two drawbacks. On the one hand, the number of parameters, $K(V + D)$, grows with the size of the analyzed corpus, therefore this method is prone to overfitting [BNJ03]. On the other hand, it only fits the parameters to a given corpus, not allowing for predicting the topic distributions of unseen documents in a predictive setting.

Latent dirichlet allocation

The above mentioned shortcomings were addressed by Blei et. al. [BNJ03] in a model called Latent dirichlet allocation (LDA). Essentially, they formulated a Bayesian version of PLSI by assuming a Dirichlet-prior for the parameters $\phi_{w,z}$ and $\theta_{z,d}$ as

- for each topic $z = 1, \ldots, K$: $\phi_{w,z} \sim \mathrm{Dir}(\boldsymbol{\beta})$ with hyperparameter $\boldsymbol{\beta}$ of length V:

$$P(\boldsymbol{\phi}_z \mid \boldsymbol{\beta}) = P(\phi_{w_1,z}, \ldots, \phi_{w_V,z} \mid \beta_1, \ldots, \beta_V) = \frac{\Gamma(\sum_{w=1}^{V} \beta_w)}{\prod_{w=1}^{V} \Gamma(\beta_w)} \prod_{w=1}^{V} \phi_{w,z}^{\beta_w - 1}, \tag{5.6}$$

  with $\phi_{w,z} > 0$ for $w = 1, \ldots, V$ and $\sum_w \beta_{w,z} = 1$.

- for each document $d = 1, \ldots, D$: $\theta_{z,d} \sim \mathrm{Dir}(\boldsymbol{\alpha})$ with hyperparameter $\boldsymbol{\alpha}$ of length K:

$$P(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) = P(\theta_{z_1,d}, \ldots, \theta_{z_K,d} \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma(\sum_{z=1}^{K} \alpha_z)}{\prod_{z=1}^{K} \Gamma(\alpha_z)} \prod_{z=1}^{K} \theta_{z,d}^{\alpha_z - 1}, \tag{5.7}$$

  with $\theta_{z,d} > 0$ for $z = 1, \ldots, K$ and $\sum_z \theta_{z,d} = 1$.

Thus, in the Bayesian framework of LDA one obtains a probability distribution over all possible values of the parameters ϕ and θ given the data A and the hyperparameters ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$)

$$P(\{\phi_{w,z}\}, \{\theta_{z,d}\} \mid \{A_{w,d}\}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{P(\{A_{w,d}\} \mid \{\phi_{w,z}\}, \{\theta_{z,d}\}) P(\{\phi_{w,z}\}, \{\theta_{z,d}\} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}{P(\{A_{w,d}\} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})}, \tag{5.8}$$

where $P(\{A_{w,d}\} \mid \{\phi_{w,z}\}, \{\theta_{z,d}\})$ is the likelihood from PLSI, Eq. (5.5), $P(\{\phi_{w,z}\}, \{\theta_{z,d}\} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_z P(\boldsymbol{\phi}_z \mid \boldsymbol{\beta}) \prod_d P(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha})$ is the prior, Eqs. (5.6,5.7), and $P(\{A_{w,d}\} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \mathrm{d}\phi \int \mathrm{d}\theta P(\{A_{w,d}\} \mid \{\phi_{w,z}\}, \{\theta_{z,d}\}) P(\{\phi_{w,z}\}, \{\theta_{z,d}\} \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ is the partition function (evidence). In LDA one finds the best estimates of ϕ and θ by maximizing the posterior, Eq. (5.8).

This problem is, in general, not analytically solvable due to the large state space and the intractability of the partition function. Yet, there exist easy-to-implement approximation schemes based on Gibbs-sampling [GS04] or Variational Bayes [BNJ03], giving rise to its wide use in numerous applications [Ble12]. The important role of LDA as the state-of-the-art method in topic modeling is further emphasized by the more than 11,000 citations[1] to its original paper [BNJ03].

Nevertheless, LDA suffers from the following drawbacks:

1. The Dirichlet-prior is simply chosen for mathematical convenience as it constitutes the conjugate prior to the multinomial distribution [GCSR03]. There is no a priori justification that this distribution should be a particularly good match for real data.

2. Due to the hyperparameters, the model is not fully non-parametric, i.e. the particular values have to be chosen in an arbitrary way or using heuristics [WMM09].

[1] According to Google scholar as of July 30, 2015.

3. The number of topics, K, is a fixed parameter in LDA and, therefore, can not be inferred. This problem is typically addressed via post-inference cross validation [BNJ03].

5.1.2. Community detection in complex networks

A network, or graph G, consists of a set of $i = 1, \ldots, M$ nodes, and a collection of E edges connecting two nodes i and j, i.e. $\{(i_1, j_1), (i_2, j_2), \ldots, (i_E, j_E)\}$. Networks have been proven to provide a useful abstraction of many complex systems in nature and society [AB02, New10], in which the nodes represent the constituents and the edges their mutual interaction, e.g. metabolic networks describing chemicals produced and consumed by chemical reactions, food webs in ecology describing species and the respective predator-prey dynamics, or social networks describing friendship between individuals. These networks can be formally described by the adjacency matrix $A = \{A_{i,j}\}$ with $i, j = 1, \ldots, M$, where $A_{i,j}$ is the number of edges between nodes i and j. In the most simple case of a binary network, $A_{i,j} = 1$ ($A_{i,j} = 0$) if there exists (does not exist) a connection between node i and j. Here we are interested in undirected (edges are symmetric, i.e. they do not have a direction) multigraphs, i.e. $A_{i,j} \in \mathbb{N}$ (there can be multiple edges between two nodes). In a more general case, edges can be directed as in the example of the food web or assigned an arbitrary weight, however, we will not consider these cases in the following.

One of the main problems in the study of complex networks is the detection of large-scale structures [For10, New11a], i.e. the identification of groups of nodes with a similar connectivity pattern. The identification of this large-scale structure is motivated by the fact that the groups i) describe the heterogeneity (i.e. the non-random structure) of the network and ii) may correspond to functional units or the intuitive notion of a community in a social network of individuals.

The problem of finding communities is intrinsically ill-posed as there exists no formal (agreed upon) definition of community structure. Often groups are defined in the sense of dense subnetworks, i.e. so-called assortative structure, in which nodes in the same group are are more connected among themselves than with the rest of the network, in which case they are called communities. As a result, there exist many different approaches [For10], where perhaps the method used most is based on the optimization of a quality function called modularity [NG04]. However, modularity (and most of the proposed methods) lack a sound statistical foundation such that it is impossible to separate actual structure from artifacts induced by statistical fluctuations or finite-size effects which constitutes a severe drawback in the interpretation of the results obtained from the respective algorithms. Note that, in the case of topic models, see Sec. 5.1.1, the very same problem motivated the probabilistic formulation of LSI in the form of generative models for texts. In the same spirit, the approach of stochastic block models proposes a generative probabilistic model for networks with modular structure such that the detection of communities corresponds to the problem of statistical inference of the parameters of the respective model.

Stochastic Block Models

Stochastic block models (SBM) are a class of generative models for networks with modular structure originally proposed in the social sciences [HLL83]. In its simplest form, each node $i = 1, \ldots, M$ is assigned to one of K groups, denoted by $\{z_i\}$ with $z_i \in [1, \ldots, K]$, and one specifies the probability for an edge between a node in group r and a node in group s by $p_{r,s}$ with $r, s = 1, \ldots, K$. This means that the SBM is parametrized by the set of group assignments $\{z_i\}$ and the $K \times K$-matrix $\{p_{r,s}\}$ in which edges between nodes in the same group are indistinguishable and are assumed to be Poisson-distributed. Given these parameters, one can write down the likelihood for an observed network with adjacency matrix A (given that there are no self-edges) [KN11]:

$$P(\{A_{i,j}\} \mid \{z_i\}, \{p_{r,s}\}) = \prod_{i<j} \frac{p_{z_i,z_j}^{A_{i,j}}}{A_{i,j}!} e^{-p_{z_i,z_j}}. \tag{5.9}$$

The problem of finding the large-scale structure is thus mapped to a problem of statistical inference by means of maximization of the likelihood with respect to the group assignments $\{z_i\}$ and the probabilities $\{p_{r,s}\}$. Even though the observed network was most probably not generated by the SBM, the inferred parameters constitute the most likely coarse-grained description of the observed network in terms of the SBM. An assortative community structure would then be revealed by nodes that are assigned to the same group s and probabilities $p_{s,s} > p_{r,s}$ for $r \neq s$. Besides the advantage of a statistical formulation, SBM are considerably more general than traditional community detection methods, since they are not restricted to assortative structures, i.e. (communities of dense links). Although this framework has been applied successfully to many real world networks, e.g. the prediction of missing or spurious edges in observed networks subject to measurement errors [GSP09], however, in practice the simple parametrization discussed above requires further modification. For example, many empirical networks show a scale-free degree distribution [BA99], in which case one has to account for the large variation in the degree of nodes within the same group by a degree-corrected SBM [KN11].

Hierarchical stochastic block model

One particular variation of the simple SBM described above was formulated by Peixoto [Pei14b, Pei15] called the hierarchical SBM (hSBM). It extends the SBM in several ways. First, in addition to incorporating the degree correction [BKN11] it also allows nodes to belong to multiple groups, i.e. it is an overlapping SBM. Second, it constructs a nested hierarchy of SBM in the following way. Once we obtain the block structure, $\{p_{r,s}\}$ of a given network with adjacency matrix $\{A_{i,j}\}$, we consider a new network with adjacency matrix $\{A'_{r,s}\}$ in which the nodes $r, s = 1, \ldots, K$ are the inferred groups of the SBM (which is smaller than the original number of nodes $i, j = 1, \ldots, M$) and the edges of the original network A are projected on the new network depending on the group membership of the nodes $i = 1, \ldots, M$ (thus the number of edges is conserved). We then infer the best block structure of the new network A' recursively, until we obtain a trivial SBM with only one block, see Fig. 5.1 for an illustration. Although this nested hierarchy of SBM require a more elaborate formulation,

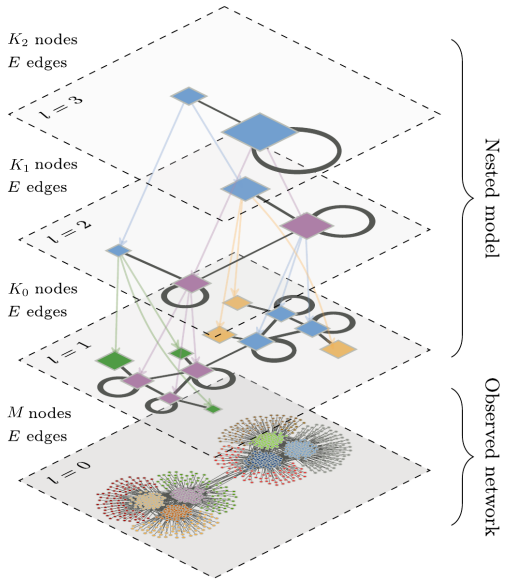

Figure 5.1.: Sketch of the hierarchical stochastic block model. Starting from the observed network ($l = 0$) with M nodes and E edges one constructs a hierarchy of L nested stochastic block models. The first level ($l = 1$) is the SBM of the observed ($l = 0$) network with K_0 groups. Each SBM at a given level l can then be considered a new network with K_{l-1} nodes and E edges for which we can determine the corresponding SBM with K_l groups until we end up with a trivial structure $K_l = 1$. In this example, the network at level $l = 3$ contains only $K_2 = 2$ nodes such that its corresponding SBM is the trivial network with $K_3 = 1$ nodes. Figure adapted from [Pei14b].

the hSBM remains tractable and yields several advantages. Not only do we obtain a description at several levels of resolution, but we can also exploit the hierarchy for a fully non-parametric Bayesian formulation. Denoting the set of parameters in the hSBM by $\{\theta\}$ we do not consider the likelihood of an observed network, $P(A \mid \{\theta\})$, but maximize the posterior distribution over the parameters given the network $P(\{\theta\} \mid A) = P(A \mid \{\theta\})P(\{\theta\})/P(A)$ by introducing a prior distribution, $P(\{\theta\})$, over the parameters. The latter is obtained in a completely non-parametric way by using the inferred SBM at an upper level as the prior information at a lower level. In principle, this maximization has to be done given the number of levels in the hierarchy, L, and the number of groups in each level of the hierarchy, K_l with $l = 1, \ldots, K$. However, we can compare hSBMs with different choices of L and K_l by employing model selection in the form of the minimum description length (MDL) [Gr7] thus finding the hSBM with the optimal values for L and K_l. In summary, the hSBM is an overlapping, degree-corrected SBM which infers a hierarchy of nested SBMs based on the statistical evidence of the given data (the network A) in a completely non-parametric way, i.e. there are no free parameters to choose beforehand.

5.2. Connecting topic models and community detection

In this section we show how the framework of community detection in complex networks, in particular SBMs, can be applied to the analysis of the topical structure in texts.

We start by noting that the word-document matrix $\{A_{w,d}\}$ in topic modeling can be represented as a bipartite network. In this representation words and documents are nodes, such that we have $M = D + V$ nodes labeled by $i = 1, \ldots, D, D + 1, \ldots, D + V$ where the first D nodes correspond to the documents d and the remaining V nodes correspond to the words w. Each occurrence of a word w in document d is considered a link between the corresponding nodes from which we obtain a multigraph (the same word w can appear multiple times in the same document d) with adjacency matrix $A_{i,j}$. Therefore, the problem of decomposing $\{A_{w,d}\}$ into the word-topic and topic-document distributions in topic modeling can be mapped to the problem of finding the block structure in the corresponding network described by the adjacency matrix $\{A_{i,j}\}$, see Fig. 5.2.

In fact, this connection has been stated conceptually [LSW+15], as well as rigorously [KN11]. In the latter case it was shown that PLSI is equivalent to the degree-corrected, overlapping SBM proposed in [KN11]: the PLSI-model with K topics is identical to the SBM with $K + D$ blocks, where each document d is assumed to belong to its own (non-overlapping) block. While LDA extends PLSI by introducing Dirichlet priors, the hSBM extends the degree-corrected, overlapping SBM by a non-parametric prior obtained from the nested hierarchy of SBMs.

Therefore, the hSBM constitutes a much more general formulation of the problem of finding the topical structure in texts solving many of the intrinsic limitations of LDA:

1. The hSBM avoids the arbitrary choice of Dirichlet priors in LDA. However, even if the empirical data (the texts at hand) had been generated by a Dirichlet distribution, the nonparametric nature of the prior in hSBM could adapt to such a situation based on the statistical evidence available. In contrast, given an arbitrary prior from the hSBM, the LDA is stuck with a Dirichlet prior. This constitutes a severe limitation as the Dirichlet distribution is unimodal and, therefore, already any non-unimodal behaviour could not be well modeled by LDA. Accordingly, the hSBM provides a much more general and versatile solution to the problem of finding topical structures in text.

2. In contrast to LDA, where the number of topics as well as the hyperparameters have to be chosen a priori or optimized by post-inference cross validation, the hSBM does not contain any free parameters. It, therefore, constitutes a much more consistent formulation in the framework of statistical inference.

3. The hSBM allows for a separate clustering of the documents and the words, respectively. While LDA assumes the same number of topics in the (symmetric) decomposition of the word-document matrix $\{A_{w,d}\}$ into the word-topic and the topic-word distributions, in hSBM the bipartite nature of the underlying network, words and documents will be assigned to disjoint blocks without imposing any symmetry. Thus, the blocks of documents can be directly used for clustering documents. In addition, the hSBM naturally infers the large-scale structure on several levels of resolution.

In a more general picture, both, topic models and community detection aim at finding large-scale structures in texts and networks, respectively. Bridging the gap between the two related, but in prac-

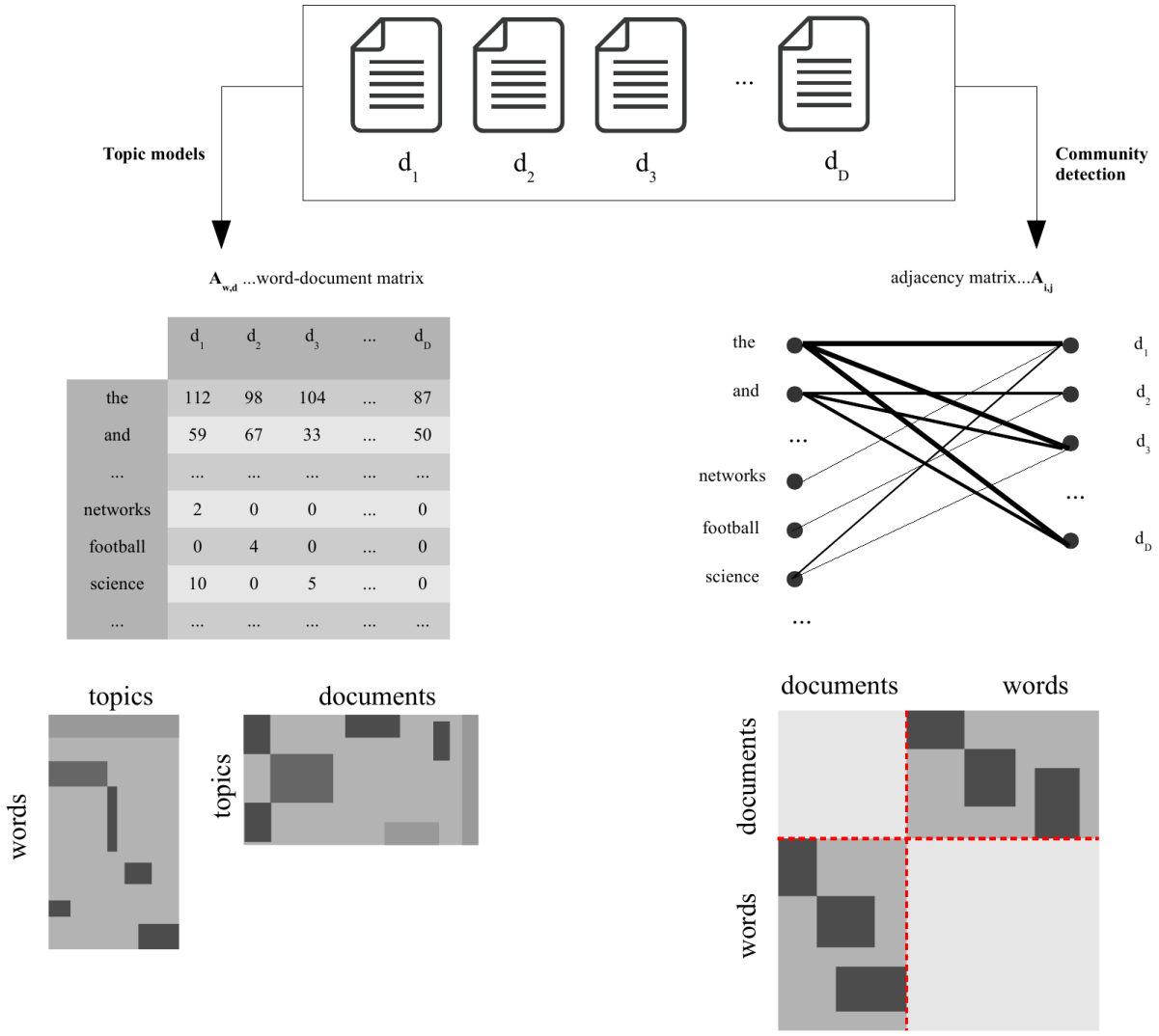

Figure 5.2.: Sketch of how to find large-scale structure in texts via topic models and community detection. The starting point are the $d_1, \ldots, d_D$ documents (box) containing written texts. In both cases the texts are treated as bag-of-words, i.e. the order of words within texts is neglected. (Left) In topic models we construct the word-document matrix $A_{w,d}$ containing the counts of each word w in each document d. This matrix is decomposed into two smaller matrices, a word-topic matrix and a topic-document matrix. (Right) In community detection we construct the bipartite multigraph in which words and documents are nodes and each occurrence of a word w in a document d is a separate edge between two nodes represented as an adjacency matrix $A_{i,j}$. The stochastic block model provides a coarse-grained description of the adjacency matrix by i) assigning nodes into groups and ii) specifying the probability of finding an edge between two groups.

tice, largely independent fields by highlighting their equivalence or similarity offers new possibilities and promises mutual benefits. On the one hand, community detection in networks has acquired a deep understanding about the underlying mechanisms of the employed algorithms. In particular, the problem of finding group structure in networks can be cast into the problem of finding the ground state of Potts spin glass-type systems, e.g. [HRN12] and references therein. These models have been well-studied in statistical and condensed matter physics and, therefore, provide a rich set of tools in the analysis of related computational problems. For example, such systems are known to exhibit phase transitions implying that even though the observed network contains structure, under certain conditions an algorithm, in principle, might not be able to infer it [DKMZ11b]. On the other hand, the problem of finding large-scale structures in texts poses new challenges for the field of community detection for which topic models provide further insight. These challenges stem from the peculiar properties exhibited by natural language often not encountered in networks. This includes, for example, the Zipfian distribution of word frequencies requiring extended formulations in terms of degree corrections, or allowing overlapping group structures. Furthermore, approaches in community detection often assume that the networks at hand are sparse, i.e. the number of edges, E, scales linearly with the number of nodes, M, as the size of the network is increased. In contrast, word-document networks do not share this property in general. In Fig. 5.3 we construct the $E(M)$-curve for the word-document network of the English Wikipedia by successively adding all $3,743,306$ articles (ordered randomly). It shows that the number of edges scales super-linearly with the number of nodes over a wide range of values for M spanning several orders of magnitude. For small and intermediate network sizes the number of documents, D, is small compared to the number of word-types, V, i.e. $D \ll V$, such that the number of nodes, M, is $M = V + D \approx V$. Since the number of edges corresponds to the number of word-tokens, i.e. $E = N$, the super-linear scaling in the $E(M)$-curve is just another representation of the sub-linear growth of the vocabulary with the number of word-tokens, $V(N)$ (Heaps' law), see Sec. 3.2, which is a direct consequence to the Zipfian word-frequency distribution. Only for extremely large collections of texts the average number of new word-types added with each new document is smaller than one. In the limit where $D \gtrsim N$, each newly added document consists effectively of one new node (the document) and (on average) a constant number of edges (the number of word-tokens contained in the document) recovering a linear scaling for $E \propto M$. In the example of Fig. 5.3 this transition can be observed for very large values of M, yet even for the complete English Wikipedia with more than 10^6 documents the $E(M)$-curve does not show a linear scaling.

5.3. Comparing LDA and hSBM

In this section, we compare LDA and hSBM in terms of which model provides a better description of a collection of texts, A. While in the previous section we explained qualitatively that hSBM provides a much more general formulation of the problem of finding topical structure, here, we provide quantitative evidence that hSBM, in general, gives better results. The Bayesian formulation of each model $\mathcal{M}$ with $\mathcal{M} \in \{\text{LDA}, \text{hSBM}\}$ as a generative probabilistic model with parameters $\theta_{\mathcal{M}}$ yields a

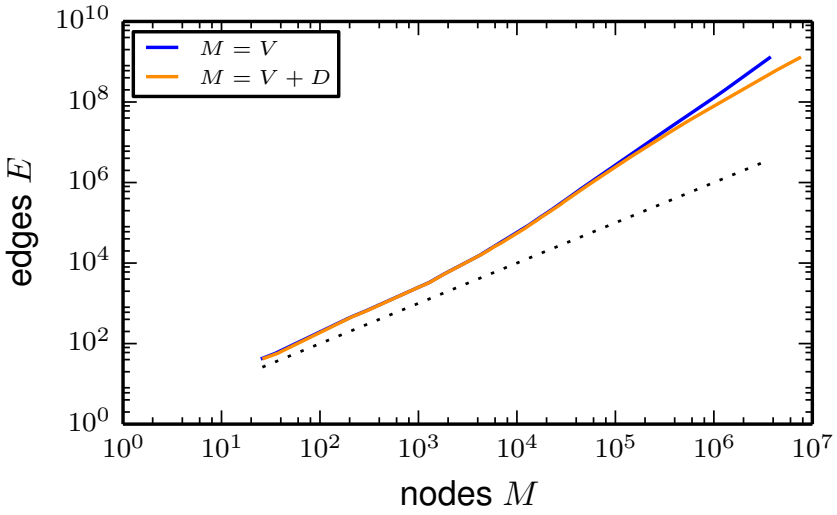

Figure 5.3.: Word-document networks are not sparse. The number of edges, E, as a function of the number of nodes, M, for the word-document network from the English Wikipedia. The network is grown by adding articles one after another in a randomly chosen order. Shown are the two cases, where i) only the V word-types are counted as nodes ($M = V$) and ii) both the word-types and the documents are counted as nodes ($M = V + D$). For comparison we show the linear relationship $E = M$ (dotted).

full posterior distribution over the parameters $P(\theta_{\mathcal{M}} \mid A)$ which allows for a comparison within the well-established framework of model selection.

In Sec. 5.3.1 we describe how to obtain the posterior distributions $P(\theta_{\mathcal{M}} \mid A)$ numerically given a collection of texts A. In Sec. 5.3.2 we discuss the problem of model selection, i.e. how to decide whether LDA or hSBM constitutes a more suitable model for a given dataset. We then apply this methodology to artificially generated corpora in Sec. 5.3.3 and to a real corpus consisting of articles from the English Wikipedia in Sec. 5.3.4.

5.3.1. Implementation

Given the finite collection of texts A (either the word-document matrix $\{A_{w,d}\}$ or the adjacency matrix $\{A_{i,j}\}$) we approximate the posterior distribution $P(\theta_{\mathcal{M}} \mid A)$ by using Monte Carlo methods. The approach in these methods is to obtain (a finite number of $s = 1, \ldots, S$) samples $\theta_{\mathcal{M}}^{(s)}$ drawn from the posterior distribution $P(\theta_{\mathcal{M}} \mid A)$ such that we get for the average of any observable $O(\theta_{\mathcal{M}})$ in the limit $S \to \infty$:

$$\langle O \rangle_S = \frac{1}{S} \sum_{s=1}^{S} O(\theta_{\mathcal{M}}^{(s)}) P(\theta_{\mathcal{M}}^{(s)} \mid A) \overset{S\to\infty}{\longrightarrow} \int \mathrm{d}\theta_{\mathcal{M}} O(\theta_{\mathcal{M}}) P(\theta_{\mathcal{M}} \mid A) = \langle O \rangle. \tag{5.10}$$

LDA

For LDA we use Mallet [McC02], a standard implementation of LDA with a Gibbs-sampling Monte Carlo method. In this approach, one assigns to each of the word-tokens one of the finite number of

$z = 1, \ldots, K$ topics such that the space of parameters becomes discrete and the set of parameters θ_{LDA} is described by $\theta_{\text{LDA}} = (\{n_{w,z}\}, \{n_{z,d}\})$, where $n_{w,z}$ is the number of word-tokens of word $w = 1, \ldots, V$ (over all documents d) that were assigned topic z and $n_{z,d}$ is the number of word-tokens in document $d = 1, \ldots, D$ (over all words w) that were assigned topic z. When evaluating summations over all possible topic assignments $(\{n_{w,z}\}, \{n_{z,d}\})$, we shorten our notation as $\sum_{\{n_{w,z}\}} \sum_{\{n_{z,d}\}} \ldots \equiv \int \mathrm{d}\theta_{\text{LDA}} \ldots$.

hSBM

For hSBM we use the implementation contained in the package graph-tool [Pei14a]. In this approach, one assigns to each half-edge (the edge incident on a given node $i = 1, \ldots, D + V$) one of the finite number of $z = 1, \ldots, K$ groups such that the space of parameters becomes discrete and the set of parameters θ_{hSBM} is described by $\theta_{\text{hSBM}} = (\{n_{i,z}\}, \{n_{z,z'}\})$, where $n_{i,z}$ is the number of half-edges incident on node i that are assigned to group z and $n_{z,z'}$ is the number edges consisting of two half-edges assigned to z and z', respectively, i.e. connecting a word group z with a document group z'. When evaluating summations over all possible group assignments $(\{n_{i,z}\}, \{n_{z,z'}\})$, we shorten our notation as $\sum_{\{n_{i,z}\}} \sum_{\{n_{z,z'}\}} \ldots \equiv \int \mathrm{d}\theta_{\text{hSBM}} \ldots$.

5.3.2. Statistical model selection

In statistical model selection the aim is to decide which model $\mathcal{M}$, parametrized by parameters $\theta_{\mathcal{M}}$, provides a better description of the data A. In our case we want to distinguish between LDA and hSBM, thus we have $\mathcal{M} \in \{\text{LDA}, \text{hSBM}\}$. Due to the fact that, typically, a model with a larger number of parameters provides a better fit of the data, one also has to take into account the complexity of each model (e.g. represented by the number of parameters) in order to avoid the problem of overfitting. This motivates the use of some principle of parsimony, also known as Occam's razor, to find the best "simple" model.

Here, we employ the principle of minimum description length (MDL) which addresses the problem of model selection in the framework of information theory by looking at the ability of a model to compress the regularities in the data [Gr7]. For probabilistic models where data and parameter space is discrete this can be quantified by the description length, Σ, i.e. the total information (in bits) needed to describe the observed data A, by

$$\Sigma(\theta_{\mathcal{M}}) = -\log P(A \mid \theta_{\mathcal{M}}) - \log P(\theta_{\mathcal{M}}) \tag{5.11}$$

where the first term corresponds to the information required to describe the data given the set of parameters of the model and the second term corresponds to the information required to describe the parameter set of the model itself. Therefore, the parameters that maximize the posterior

$$\hat{\theta}_{\mathcal{M}} = \arg\max_{\theta_{\mathcal{M}}} P(\hat{\theta}_{\mathcal{M}} \mid A). \tag{5.12}$$

yield the set of parameters that compress the data the most such that the MDL of the model $\mathcal{M}$, $\Sigma_{\mathcal{M}}$, is given by

$$\Sigma_{\mathcal{M}} = \min_{\theta_{\mathcal{M}}} \Sigma(\theta_{\mathcal{M}}) = -\log P(A \mid \hat{\theta}_{\mathcal{M}}) - \log P(\hat{\theta}_{\mathcal{M}}) \tag{5.13}$$

We note, that the MDL provides a bias towards LDA. Unlike the hSBM which is fully non-parametric, LDA requires the a priori specification of the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This description is beyond the generative process of LDA, thus, this information is not contained in the description length (and the MDL) of LDA.

In practice, we find $\hat{\theta}_{\mathcal{M}}$ by a simple annealing procedure. We first let the respective Monte Carlo algorithm equilibrate (burn-in period) at a finite inverse temperature. After that, we switch discontinuously to an infinite inverse temperature such that the Monte Carlo algorithm yields successive samples $s = 1, \ldots, S$ of the parameters $\theta_{\mathcal{M}}^{(s)}$ with $P(\theta_{\mathcal{M}}^{(s+1)} \mid A) \geq P(\theta_{\mathcal{M}}^{(s)} \mid A)$. Therefore, in the limit $S \to \infty$, $\theta_{\mathcal{M}}^{(S)}$ provide the parameters where $P(\theta_{\mathcal{M}} \mid A)$ shows local maxima. Repeating this procedure with different initial conditions and selecting the best parameters among them provides an estimate for $\hat{\theta}_{\mathcal{M}}$.

5.3.3. Application: Artificial texts

In this section we consider artificial texts sampled from a given generative process which we then infer via LDA and hSBM. These toy corpora greatly simplify the problem of inference by the fact that, e.g. the LDA-hyperparameters, the number of topics, or the degree of structure, is known, and thus avoid the problems and intricacies encountered in the identification of large-scale structure in real texts.

Texts drawn from LDA

In this example we consider artificial texts sampled from the generative process defined by LDA in which the LDA-hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as well as the number of topics are fixed. Even though such corpora are highly unrealistic, this problem serves as a consistency check for inference with LDA by assuming that the assumption of Dirichlet-priors is actually true. Note that in the comparison (model selection) with hSBM, this particular problem gives an advantage to LDA since (only in this particular case) the generative process of the corpora and the generative process underlying the inference are identical. Considering topic models and community detection in the framework of dimensionality reduction, i.e. to find large-scale structure (see Sec. 5.1), we are interested in the case $K < N, D$, i.e. the number of topics is smaller than the number of words and documents. Our findings in Fig. 5.3 show that for small and intermediate (large) collections of real texts we have $D < N$ ($D > N$). Therefore, we consider two cases, i.e. i) $K \ll N \ll D$ ($K = 10$, $N = 100$, $D = 1000$) and ii) $K \ll D \ll N$ ($K = 10$, $D = 100$, $N = 1000$), keeping the number of word-tokens in each document (the text length) fixed ($n_d = 100$). For sake of simplicity we assume flat Dirichlet-priors specified by scalar LDA-hyperparameters α and β, i.e. $\boldsymbol{\alpha} = (\alpha_1 = \alpha, \ldots, \alpha_K = \alpha)$ and $\boldsymbol{\beta} = (\beta_1 = \beta, \ldots, \beta_N = \beta)$, fixing $\beta = 0.1$ and varying α such that for low (high) values of α we get a single topic (all topics with

equal weight) in a single document.

In Fig. 5.4 we compare the MDL for LDA and hSBM for numerically generated corpora. Even though the corpora were sampled from and inferred with the (same) generative process of LDA, hSBM provides a smaller MDL and thus constitutes a better model over a wide range of parameters α. This seemingly contradiction can be understood by looking closer at the corpora sampled from LDA: the word-topic distribution $\phi_{w,z}$ is obtained from K samples of the N-dimensional Dirichlet distributions with hyperparameter $\boldsymbol{\beta}$ and the topic-document distribution $\theta_{z,d}$ is obtained from D samples of the K-dimensional Dirichlet distribution with hyperparameter $\boldsymbol{\alpha}$. In the cases considered here (which we argued are the practical cases of interest), $K \ll N$, such that the Dirichlet distribution over the words is always under-sampled even when looking at arbitrarily large corpora. This means that, from the data it is impossible to learn that the data is generated from this distribution and there are possibly other (simpler) ones that describe the observations equally well. The non-parametric formulation of hSBM yields a better description of the artificial corpora because it does not explicitly assume Dirichlet-priors. This finding emphasizes the intrinsic problems underlying the assumption of Dirichlet-priors in LDA as well as the advantage of the non-parametric formulation of hSBM. Even if we construct artificial texts in which this assumption is valid, LDA does not necessarily offer the best way in which the structure in these texts can be compressed.

Texts with a planted structure

In this example we consider artificial texts with a planted structure in analogy to the benchmark graphs used in the comparison of community detection algorithms by Ref. [LFR08]. The aim is to generate corpora in which we fix the number of topics and control the degree of structure (which we will call assortativity c) by varying a single parameter c from $c = 0$ (no structure) to $c = 1$ (full

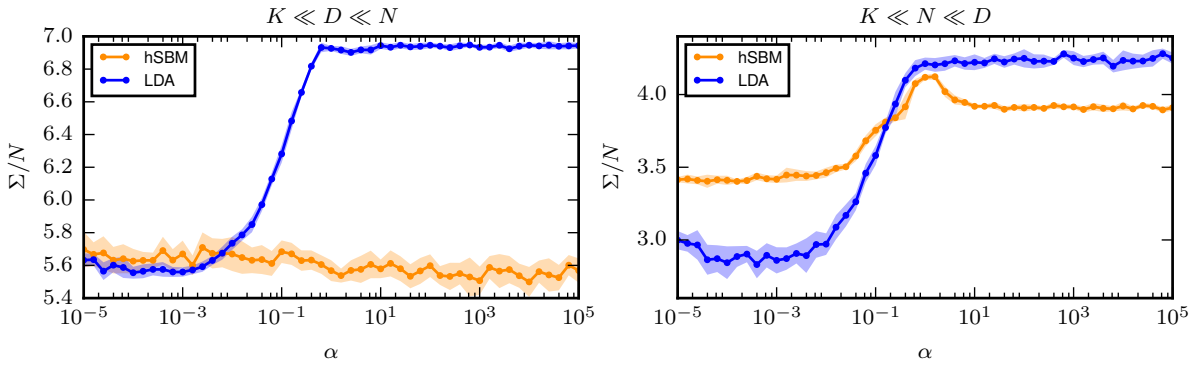

Figure 5.4.: Model selection for texts drawn from LDA. Inferred values of the Σ for LDA and hSBM for artificial texts drawn from the generative process of LDA with $\beta = 0.1$ and varying α in the two regimes $K \ll N \ll D$ (left) and $K \ll D \ll N$ (right), see main text for details. In the inference with LDA we fix the number of topics $K_{\mathrm{LDA}} = K = 10$ and use as hyperparameters $\alpha_{\mathrm{LDA}} = \alpha$ and $\beta_{\mathrm{LDA}} = \beta = 0.1$ the same parameters that were used to generate the corresponding corpus. In the inference with hSBM we fix the number of groups as $K_{\mathrm{hSBM}} = K + D$ (each document is assigned to a separate group) in order to allow for a meaningful comparison between the two models, see Sec. 5.2. The curves (shaded region) show the average (standard deviation) from 10 realizations.

structure).

In our approach, we specify the size of the topics (in word-tokens) by the distribution p_z and a global word frequency distribution F_w (e.g. Zipf's law) independent of the topics. The artificial corpus is then defined by a generative process in which i) each document contains exactly one topic z_d drawn from p_z, and ii) each word-token in document d is drawn from a word-topic distribution $\phi^{(c)}_{w,z=z_d}$. Assigning each word-type w to a single topic z_w we assume that $\phi^{(c)}_{w,z}$ is a superposition of a structure-term and a noise-term

$$\phi^{(c)}_{w,z} = cF^{(\text{struct})}_w \delta_{z_w,w} + (1-c)F^{(\text{noise})}_w. \tag{5.14}$$

Imposing the global word-frequency distribution, i.e. $F_w \equiv \sum_z \phi^{(c)}_{w,z}$, and satisfying the normalization $\sum_w \phi^{(c)}_{w,z=z_d} = 1$, we get

$$\phi^{(c)}_{w,z} = c\frac{F_w}{p_z}\delta_{z_w,w} + (1-c)F_w \tag{5.15}$$

with the constraint $\sum_{w:z_w=z} F_w = p_z$. In Fig. 5.5 we give an example of the dependence of $\phi^{(c)}_{w,z=z_d}$ on the assortativity parameter c. For an arbitrary F_w and a word-topic assignments z_w, it is hard to exactly fulfill the condition $\sum_{w:z_w=z} F_w = p_z$. In practice, for each word-type w we randomly draw a topic assignment $z_w = z$ from p_z and by approximating $p_z \approx \tilde{p}_z \equiv \sum_{w:z_w=z} F_w$ by the empirically drawn z_w fulfills the condition $\sum_{w:z_w=z} F_w = p_z$ by construction. In the following we restrict ourselves to the case where each topic is equally probable, i.e. $p_z = 1/K$. We consider two different cases, where

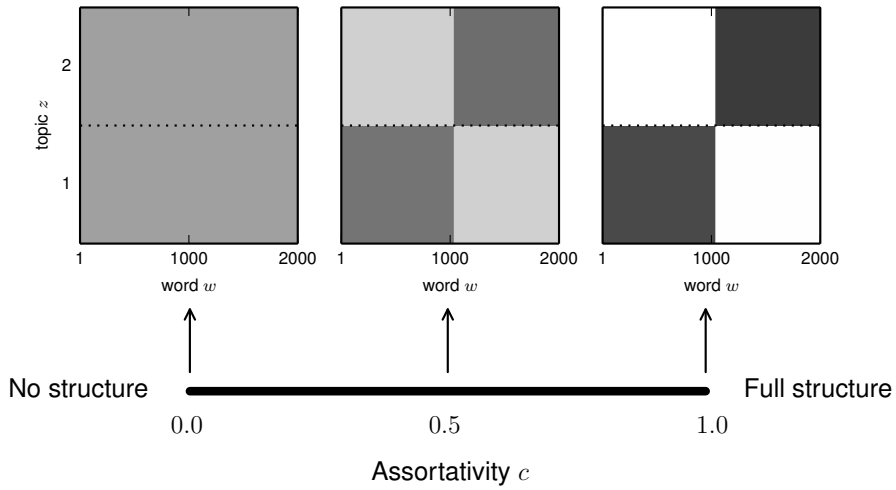

Figure 5.5.: Illustration of the planted structure model. Frequency of word w given the topic z, $\phi^{(c)}_{w,z}$ in Eq.(5.15), shown for the values $c \in \{0.0, 0.5, 1.0\}$ for the case of $K = 2$ topics, i.e. $z = 1, 2$ and $N = 2000$ word-types, i.e. $w = 1, 2, \ldots, 2000$, with equiprobable topic sizes, i.e. $p_z = 1/K$, and a flat word frequency distribution, i.e. $F_w = 1/N$. Darker colors indicate a larger frequency.

the word-frequency distribution, F_w, is i) flat, i.e. $F_w = 1/V$; and ii) follows Zipf's law, i.e. $F_r \propto r^{-\gamma}$ with $\gamma = 1.5$ by assigning a rank $r = r(w) = 1, \ldots, V$ to each word w.

In Fig. 5.6 we compare the MDL of LDA and hSBM for corpora ($K = 10$, $N = 10^5$, $D = 10^3$, $n_d = 200$) with assortatitivity $c \in [0, 1]$. We observe in both cases ('flat' and 'zipf') that the hSBM constitutes a much better model for the whole range of parameters c as it shows systematically lower values for the MDL. Looking at the individual values for the MDL as a function of c for the flat case in Fig. 5.6 we observe another striking feature. For large values of c, the MDL is decreasing with c indicating that the we can obtain a better compression (and thus a better inference result) the more structure there is in the data. However, there exists a finite $c^* > 0$ such that the MDL is constant for $c < c^*$. This is remarkable insofar as for $0 < c < c^*$ the data is not completely random containing a finite degree of structure but the inference results obtained from both LDA and hSBM yield the same description length as in the completely unstructured data ($c = 0$). In other words, the most likely inference result found in this regime is indistinguishable from the random assignment of word-tokens to groups. Thus we conjecture that at the parameter $c = c^*$ a phase transition from an undetectable to a detectable phase takes place, a behaviour that has recently been explored in simpler versions of the SBM in Ref. [DKMZ11b]. Note that, in the latter case, the authors prove the existence of such a phase transition for a belief-propagation algorithm that they show provides the asymptotically correct inference results in the case of SBM. Our statement is milder in the sense that, here, we provide a simple example in which both LDA and hSBM exhibit an undetectable phase. That is, even though by construction the corpus contains a finite degree of structure, the most likely inference result provides no more information (as measured by the description length) than in a completely random corpus. Furthermore, comparing the flat and the Zipfian word-frequency distribution in Fig. 5.6, we observe

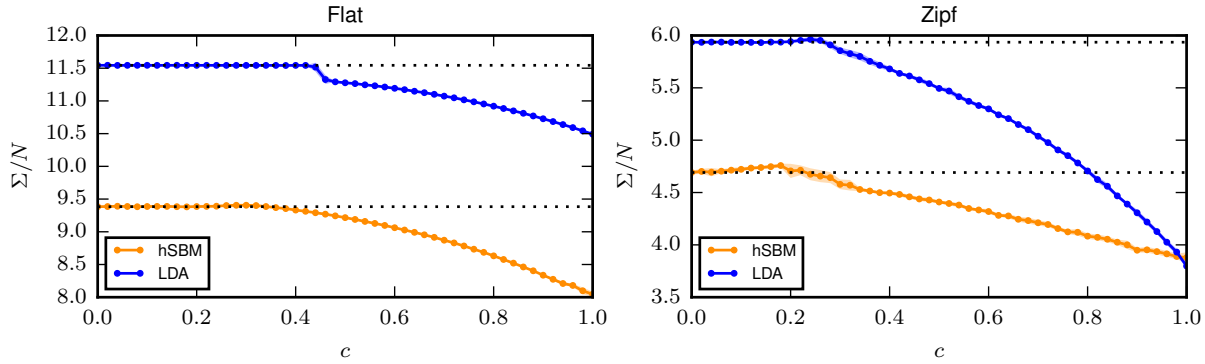

Figure 5.6.: Model selection for texts with a planted structure. Inferred values of the Σ for LDA and hSBM for artificial texts varying the degree of structure, c, for a flat (left) and Zipfian (right) word frequency distribution, see main text for details. In the inference with LDA we fix the number of topics $K_{\mathrm{LDA}} = K = 10$ and use Mallet's heuristic optimization procedure for the hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ described in [WMM09]. In the inference with hSBM we fix the number of groups as $K_{\mathrm{hSBM}} = K + D$ (each document is assigned to a separate group) in order to allow for a meaningful comparison between the two models, see Sec. 5.2. For visual comparison we show the MDL-values for the case $c = 0$ (dotted lines). The curves (shaded region) show the average (standard deviation) from 10 realizations.

that the value of c^* becomes smaller for the Zipfian case, i.e. the region where we are unable to infer any structure shrinks. This is particularly interesting in view of the ubiquitousness of Zipfian word-frequency distribution in natural languages, as it makes inference easier compared to word-frequency distributions that are flat.

5.3.4. Application: Real texts

In this section we consider real texts from the English Wikipedia restricting ourselves to articles that belong to the category "Scientific Disciplines". In detail, our corpus consists of 4219 articles taken from 100 different randomly selected sub-categories.

In Fig. 5.3.4 we compare the MDL for LDA and hSBM. Since in this case the "true" number of topics is unknown, we vary the number of topics, K_{LDA}, used in the inference with LDA. We find a minimum in the MDL for $200 < K_{\mathrm{LDA}} < 2000$ indicating the optimal choice for the number of topics in LDA. For hSBM the optimal number of topics is automatically determined ($K_{\mathrm{hSBM}} = 276$ groups) yielding a much lower MDL, and thus a better description of the data, than all LDA-models. Note that the number of topics in LDA and the number of groups in hSBM can not be compared directly. In LDA, the partition of words and documents into topics is symmetric since the inferred word-topic and topic-document distributions are based on the same topics. In contrast, hSBM partitions all nodes (word-types and documents) of the word-document network. Due to the bipartite nature of the network, words and documents are automatically assigned into different groups such that the total number of groups consists of groups for words and groups for documents, which are not necessarily of the same magnitude. Furthermore, we remark that due to computational limitations, we only considered a non-overlapping version of the hSBM. However, every overlapping hSBM can be mapped

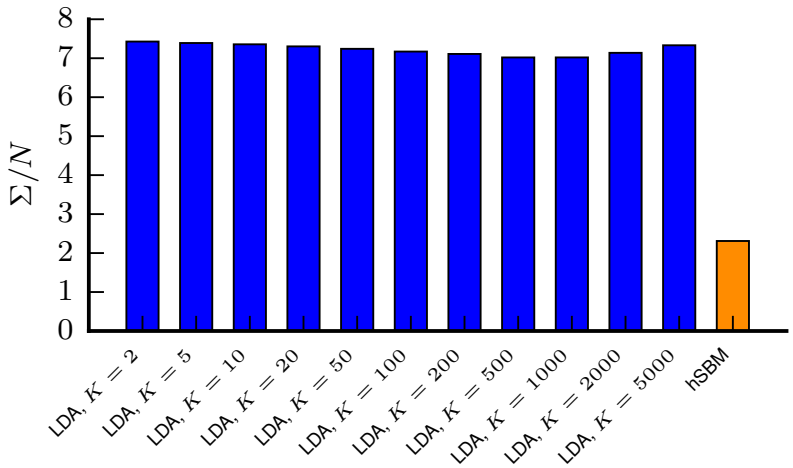

Figure 5.7.: Model selection for LDA and hSBM in real texts. Inferred values of Σ for i) LDA-models with different number of topics (blue) and ii) hSBM, where the optimal number of groups is automatically determined (here $K_{\mathrm{hSBM}} = 276$) for a subset of articles from the English Wikipedia contained in the category "Scientific Disciplines" with $V = 80,244$, $D = 4,219$, and $N = 3,381,465$.

to an equivalent non-overlapping hSBM [Pei15], where the MDL-principle of model selection provides an answer as to which model is more suited to describe the data at hand. In fact, it was shown in [Pei15] that many real networks are better described by a non-overlapping hSBM. Here, our main point is to illustrate the usefulness of the framework of the stochastic block models and, more generally, the methods of community detection from complex networks, in describing topicality in texts.

In Fig. 5.8 we show the inferred result of hSBM applied to the word-document network of the English Wikipedia. It provides a visual intuition to the result obtained from hSBM and emphasizes the usefulness of the hierarchy in the description of the large-scale structure. At the highest level of the hierarchy (middle), the nodes of the network are split into two large groups, i.e. words (bottom) and documents (top), which are then further split into smaller and smaller groups. Therefore, we obtain a coarse-grained description of the word-document network on several different scales of granularity. Examining individual groups, we find that the inferred partitioning of words and documents corresponds to our intuition of semantic categories. For example, on the level of documents, one group can be identified as containing articles about electromagnetism (e.g. "Lennard-Jones potential" or "Maxwell's equations"), whereas another contains articles related to the concept of color in computer graphics (e.g. "Color image" or "Color gradient"). On the level of words, we identify similar groups of word-types associated to these concepts (e.g. "electromagnetic" and "wave" for electromagnetism and "image" and "graphics" referring to the concept of computer graphics). Furthermore, we see an additional type of group containing function words (e.g. "and", "in", or "on") which are not associated to any specific context. The reason for the appearance of the latter groups is that we employed a non-overlapping version of the hSBM, such that words (or documents) that would be contained in several groups in the overlapping hSBM are simply inferred as a separate group.

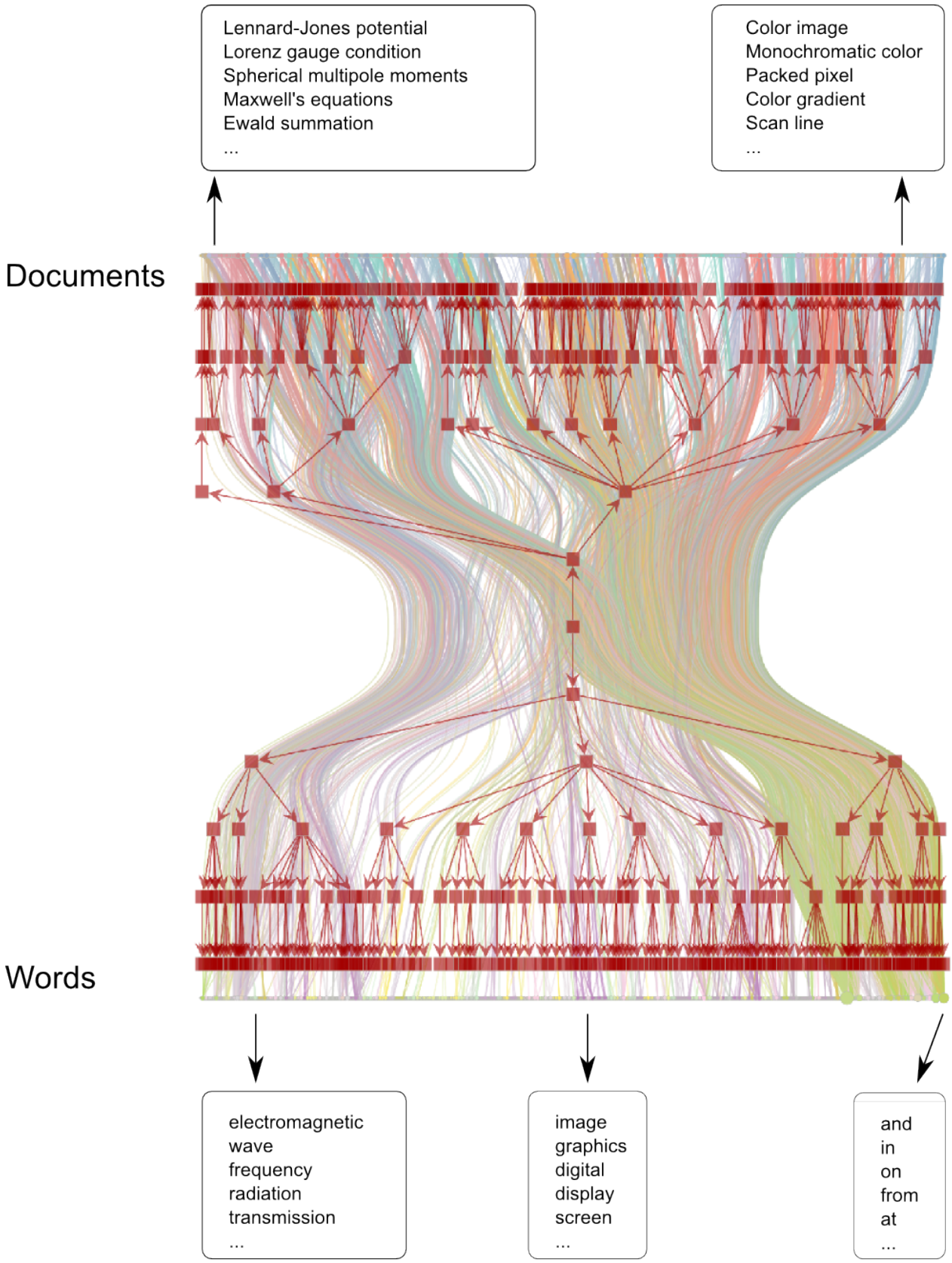

Figure 5.8.: Visualization of the inference result of hSBM for texts from the English Wikiepdia. Word-document network (same data as in Fig. 5.3.4) where i) nodes correspond to word-types (bottom) and documents (top) and ii) edges correspond to each word-token appearing in a document. Nodes of the same color belong to the same inferred groups on the lowest level of the hierarchy. The boxes show examples of assignment of nodes to groups, i.e. 5 random articles (the 5 most common word-types) in groups of documents (words). Rectangular nodes and arrows show group structure on different levels of the hierarchy.

6. Variability in time

In this chapter we study the variability of language over time also known as diachronic language change [WLH68, Cro00]. While in Ch. 3 we showed that natural language exhibits a remarkable degree of universal structure in terms of the temporal stability of certain macroscopic observables, e.g. the double power law rank-frequency distribution, languages are constantly subject to transformation processes. The study of language change concerns many different aspects, e.g. the introduction of new words, the change in meaning, or the change in pronunciation (sound shifts) [Ait01]. Here, we focus on the change in the frequency of usage of fixed word forms (i.e. words with a given spelling). The magnitude of recently available databases, in particular the Google-ngram data, both, in terms of the size in number of words, but also in terms of the temporal resolution, enables us to trace the changes in the usage of language quantitatively on historical timescales.

In this we take two different approaches. In Sec. 6.1 we will assess how the vocabulary of a language, i.e. the ensemble of words as measured by their frequency of usage in a given year in the Google-ngram corpus, is evolving over time [GA13, GFCA15]. This requires the application of statistical methods in order to obtain reliable quantitative measures. This becomes especially important when considering the fat-tailed character of the underlying distribution of word frequencies. In Sec. 6.2 we are interested in the dynamical behaviour of individual word forms asking how new words (or linguistic innovations more generally) are spreading through a community of speakers [GGMA14]. In the framework of statistical physics, the aim is to combine microscopic models describing the process of word usage on the level of individuals (where the interaction takes place, e.g., on a complex network) with empirical data usually only available on the macroscopic level, e.g. the fraction of the population that has adopted a given linguistic innovation.

6.1. Change in the vocabulary of a language

In this section, we quantify the change in the vocabulary of a language on historical timescales. In Sec. 6.1.1 we investigate the change in the composition of the core vocabulary identified in the statistical analysis of word frequencies in Ch. 3. In a more general approach in Sec. 6.1.2 we compare the word-frequency distributions of the Google-ngram database from different years by applying the framework of the spectrum of divergences, $\tilde{D}_\alpha$, developed in Sec. 4.3.

6.1.1. Decay of the core vocabulary

The analysis of the distribution of word frequencies in Sec. 3.1 in the form of a generalized Zipf's law, $f_{\rm dp}(r;\gamma,b)$ in Eq. (3.16), with double scaling has been shown to give a good account for all databases and all years in the Google-ngram database with the same fixed two parameters $b^* = 7,873$ and $\gamma^* = 1.77$ in the case of English. In Sec. 3.2.2 we provided an interpretation of the transition between the two scalings in terms of the existence of a core vocabulary of finite and constant size $V_c^{\rm max} = b^* = 7,873$.

A natural question is, therefore, what actually changes in historical time scales? Considering two different databases (say two different years), our description does not consider any differences in the actual composition of the database. Even if the value of $V_c^{\rm max}$ remains constant this does not mean that the *same* words are observed for all years. From the point of view of our generative model in Sec. 3.2.2, the main change a word can experience is to enter or to leave the group of core words. For instance, comparing the decades $1891-1900$ and $1991-2000$, the most frequent words which left the core vocabulary were *majesty*, *doubtless*, *furnished*, *monsieur*, *napoleon*, and *hitherto*, while the ones which entered were *cultural*, *context*, *technology*, *programs*, *environmental*, and *computer* [1].

In order to quantify this effect, we investigate the replacement of words from the core vocabulary in the yearly databases $y(t)$ in the time $t \in [1805, 2000]$ in Fig. 6.1. We calculate the fraction $F(t,\Delta t)$ of core words (i.e. with rank $r < b^* = 7873$, fixed for all t) from $y(t)$ that remain in the set of core words in $y(t+\Delta t)$. Figure 6.1(a) shows that all curves can be qualitatively described by an exponential

[1] These examples are the 6 most frequent words which belonged to the core vocabulary (i.e., $r < b^* = 7873$) in every single year in one decade and in none of the years in the other decade (ordered by the average frequency in the decade which they belonged to the core vocabulary)

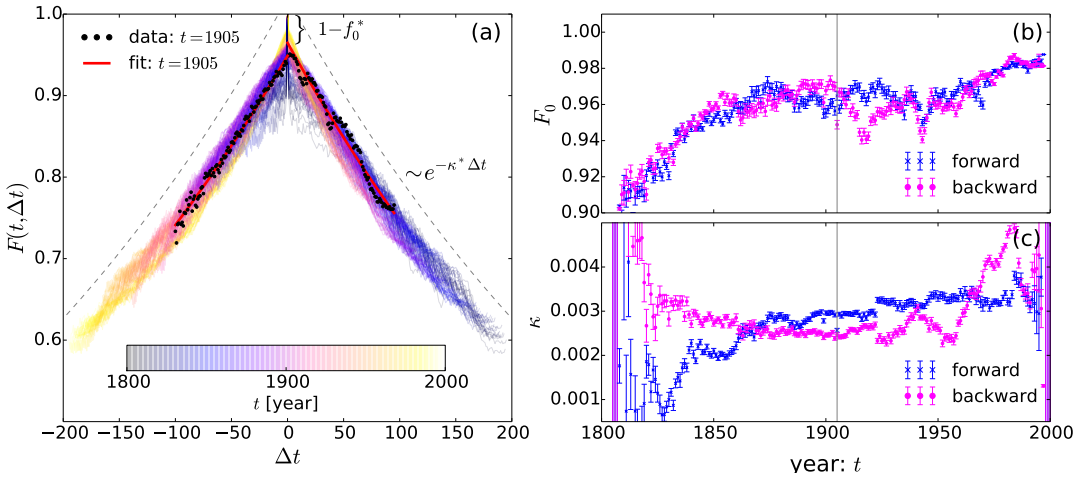

Figure 6.1.: Historical change in the composition of core words in the English vocabulary. a) fraction $F(t,\Delta t)$ of core words in $y(t)$ that remain among this set in $y(t+\Delta t)$ for $t \in [1805, 2000]$ (pale colors) and in particular for $t = 1905$ (black dots) with the corresponding exponential fit (red line). b+c) Parameters F_0 and κ in the exponential decay Eq. (6.1) of the curves in a) obtained through least-square fits. Forward (backward) decay refers to $\Delta t > 0$ ($\Delta t < 0$).

decay

$$F(t, \Delta t) = F_0 e^{-\kappa|\Delta t|}, \tag{6.1}$$

independent of whether forward ($\Delta t > 0$) or backward time ($\Delta t < 0$) was considered. This is further supported in Fig. 6.1(b-c), where the parameters F_0 and κ obtained numerically from a least-square fit [HTF09] of Eq. (6.1) for all curves $F(t, \Delta t)$ with $t \in [1805, 2000]$ are presented. In order to avoid biases due to different number of points in the fit, for each t we performed a fit with the same number of points $\min\{2000 - t, t - 1805\}$ forwards and backwards in time. On closer inspection, two features connected to the interpretation of the parameters F_0 and κ deserve a more careful discussion. The parameter $F_0 < 1$ represents the discontinuous change of core words in two subsequent years. It strongly depends on the different selection of books in the construction of the respective databases and can be attributed to the finite size of the database, which leads to a wrong estimation of the "true" core words. Consistently with this interpretation, Fig. 6.1(b) shows that F_0 grows over time, due to the fact that database size increases leading to a better sampling of words. Nevertheless, a value of $F_0 \approx 0.98$ indicates that this is still far from being negligible (e.g., for $V_c^{\max} = 7,873$ this means that around 150 words of the set of core words will be different due to finite sampling). In contrast, the decay rate κ describes the continuous replacement of core words over time with a rate of $\kappa V_c^{\max} \approx 30$ words per year. The most intriguing observation in Fig. 6.1(c) is that this change experiences an acceleration over time as κ grows by more than 50% from 1805 to 2000.

Finally, it is worth to compare these numbers with other recent studies on historical changes in language usage which reported half-lives for: i) the regularization of verbs (750 to 10 000 years) [LMJ+07], and ii) a fundamental vocabulary of 200 words (300 to 38 000 years) [PAM07]. The most intriguing observation in Fig. 6.1(c) is the approximately linear increase of the decay rate over time as κ grows by more than 50% from 1805 to 2000. This could be interpreted as a confirmation of the overall acceleration of language change and society in general, as propagated in Ref. [MSA+11].

6.1.2. Measuring language change by $\tilde{D}_\alpha$

In this section, we apply the framework of the spectrum of divergences (see Sec. 4.3) in the quantification of language change. We calculate the normalized spectrum $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$, Eq. (4.33), between pairs of word-frequency distributions of the Google-ngram database from different years $t_1 \neq t_2$. In this approach, we are not only interested in following the temporal behaviour by comparing two or more divergences for fixed α, e.g. $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ and $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_3})$ with $t_1 \neq t_3$, but also how different choices of α yield different results for a fixed pair of years (t_1, t_2) by comparing, e.g. $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ and $\tilde{D}_{\alpha'}(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ with $\alpha \neq \alpha'$. As argued in Appendix D, $\tilde{D}_\alpha(\boldsymbol{p}, \boldsymbol{q})$ is meaningful even if the sequences used to estimate $\boldsymbol{p}$ and $\boldsymbol{q}$ have different sizes $N_p \neq N_q$. We summarize our results in Fig. 6.2, from which different conclusions can be drawn:

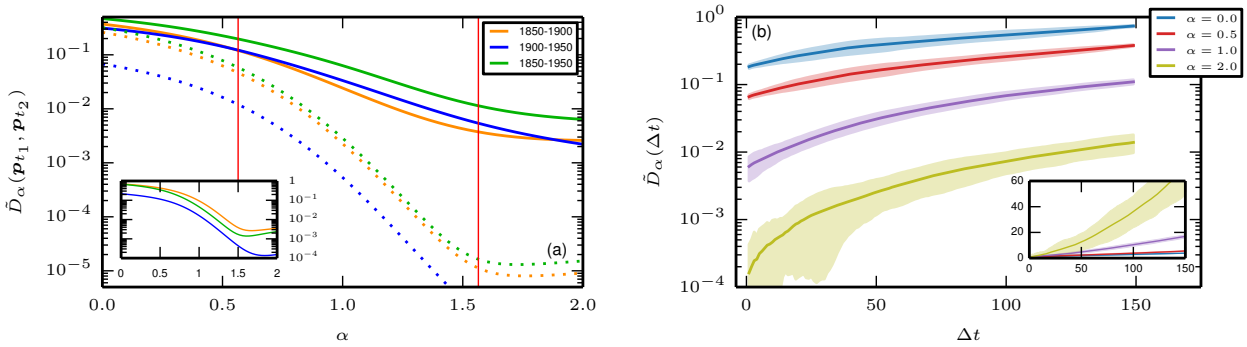

Figure 6.2.: Measuring change in the usage of language on historical time scales. (a) $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ as a function of α for pairs of word frequency distributions of the Google-ngram database obtained from the yearly corpora t_1 and t_2 with $(t_1, t_2) \in \{(1850, 1900), (1900, 1950), (1850, 1950)\}$ (solid lines). The dotted lines with the same colors show the results of a null model in which samples of the same size of the ones in t_1 and t_2 are randomly drawn from the *same* distribution (obtained combining the corpora in t_1 and t_2). The vertical lines show the three regimes $\alpha < 1/\gamma$, $1/\gamma < \alpha < 1+1/\gamma$, and $\alpha > 1+1/\gamma$ in the convergence of $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ with N (see Sec. 4.3.2), obtained using $\gamma = 1.77$ (see Sec. 3.1). Inset: ratio $\tilde{D}_\alpha(\boldsymbol{p}_{t_{12}}, \boldsymbol{p}_{t_{12}})/\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$. (b) Average divergence as a function of $\Delta t \equiv |t_2 - t_1|$, calculated as $\tilde{D}_\alpha(\Delta t) = \frac{1}{N_{\Delta t}} \sum_{t_1=1805}^{2000-\Delta t} \tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_1+\Delta t})$ for four different α (solid lines). Shaded areas represent the standard deviation associated to the average $\tilde{D}_\alpha(\Delta t)$. Inset: $\tilde{D}_\alpha(\Delta t)/\tilde{D}_\alpha(\Delta t = 1)$.

Temporal change

The change of English from 1850 to 1950 was larger than the change form 1850 to 1900 and from 1900 to 1950, as seen from the fact that the curve of $\tilde{D}_\alpha(\boldsymbol{p}_{1850}, \boldsymbol{p}_{1950})$ in Fig. 6.2(a) lies above the two other curves for all α. This intuitive result (evolutionary dynamics show no recurrences) confirms that the divergence spectrum $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ leads to a meaningful quantification of language change. The average dependency of $\tilde{D}_\alpha(\boldsymbol{p}_{1850}, \boldsymbol{p}_{1950})$ on $\Delta t = |t_2 - t_1|$, shown in Fig. 6.2(b), can be thus used as a quantification of the speed of language change.

Dependence on α

All observed divergences $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ decay with α [e.g., the three curves in Fig. 6.2(a)]. As discussed in Sec. 4.3.2, this shows that for words with a high (low) frequency the distributions are more (less) similar and thus the change is slower (faster). This result is consistent with reports that frequent words tend to be more stable on historical time scales [PAM07, LMJ+07]. This dependence on α is essential when comparing the change $1850 \mapsto 1900$ to the change $1900 \mapsto 1950$ [Fig. 6.2(a)]. While the earlier change was smaller if counted on a token basis, $\tilde{D}_{\alpha=1}(\boldsymbol{p}_{1850}, \boldsymbol{p}_{1900}) < \tilde{D}_{\alpha=1}(\boldsymbol{p}_{1900}, \boldsymbol{p}_{1950})$, it becomes larger if one focuses on the more frequent words $[\tilde{D}_{\alpha=2}(\boldsymbol{p}_{1850}, \boldsymbol{p}_{1900}) > \tilde{D}_{\alpha=2}(\boldsymbol{p}_{1900}, \boldsymbol{p}_{1950})]$.

Role of finite-size scalings

Our finding in Sec. 4.3 that the scaling of the bias and fluctuations in $\tilde{D}_\alpha$ with sample size N depend on α allows for a deeper understanding of the $\tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$ measurements discussed above. The expected

$\tilde{D}_\alpha$'s for random sampling of the same distribution [null model shown as dashed line in Fig. 6.2(a)] are of the same order as the empirical distance for small α (i.e. $\tilde{D}_\alpha(\boldsymbol{p}_{t_{12}}, \boldsymbol{p}_{t_{12}}) \approx \tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$) and it is only for $\alpha > 1$ that the null model divergence becomes negligible compared to the empirical divergence (i.e. $\tilde{D}_\alpha(\boldsymbol{p}_{t_{12}}, \boldsymbol{p}_{t_{12}}) \ll \tilde{D}_\alpha(\boldsymbol{p}_{t_1}, \boldsymbol{p}_{t_2})$). This implies that even though the size of the individual corpora is of the order of $N \approx 10^9$ word-tokens, the empirically measured $\tilde{D}_\alpha$ is still strongly influenced by finite-size effects over a wide range of values for α, in agreement with our analysis in Sec. 4.3.2. It also emphasizes that $\tilde{D}_{\alpha=2}$ offers a pragmatic choice in reducing such finite-size effects when the exponent in the power-law distribution is not known. This conclusion is further corroborated in the analysis of the dependence of $\tilde{D}_\alpha$ with Δt [Fig. 6.2(b)]. For small α, $\tilde{D}_\alpha$ does not converge to zero for $\Delta t \to 0$, but instead it seems to saturate, i.e. $\tilde{D}_\alpha(\Delta t \to 0) \approx \tilde{D}_\alpha(\Delta t = 1) = \epsilon_\alpha > 0$. The value ϵ_α is of the same order of magnitude of the expected bias [e.g., shown as dashed line in Fig. 6.2(a)] and, for small α, still of the same order of magnitude of the distance $\tilde{D}_\alpha(\Delta t = 100)$ between two corpora separated by 100 years. For small α and Δt, it is thus difficult to distinguish between finite-size effects (ϵ_α) and actual language change. Results for $\alpha = 2$ show the largest relative variation with Δt [see Inset of Fig. 6.2(b)] and are therefore more suited to discriminate and quantify language change over time.

6.2. Innovation of new words

In this section we focus on the temporal change in the frequency of usage of individual words in particular the innovation and spreading of new words.

It is well accepted that adoption of innovations are described by S-curves which often amounts to the *qualitative* observation that the change starts slowly, accelerates, and ends slowly. Linguists generally accept that *"the progress of language change through a community follows a lawful course, an S-curve from minority to majority to totality."* [WLH68], see Ref. [BC12] for a recent survey of examples in different linguistic domains. *Quantitative* analysis are rare and extremely limited by the quality of the linguistic data, which in the best cases have *"up to a dozen points for a single change"* [BC12]. Going beyond qualitative observation is essential to address questions like:

(i) Are all changes following S-curves?

(ii) Are all S-curves the same (e.g., universal after proper re-scaling)?

(iii) How much information on the process of change can be extracted from S-curves?

(iv) Based on S-curves, can we identify signatures of endogenous and exogenous factors responsible for the change?

Large records of written text available for investigation provide a new opportunity to quantitatively study these questions in language change [MSA+11, LMA+12]. In Fig. 6.3 we show the adoption curves of three linguistic innovations for which words competing for the same meaning can be identified. Our methodology is not restricted to such simple examples of vocabulary replacement and can be applied to other examples of language change and S-curves more generally. Here we restrict ourselves

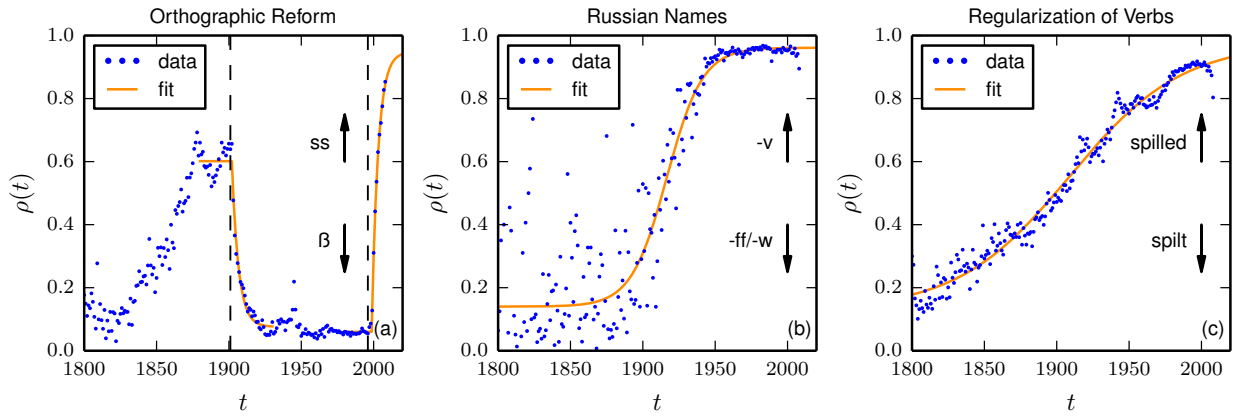

Figure 6.3.: Examples of linguistic changes showing different adoption curves. We estimate the fraction of adopters $\rho(t)$ by the relative frequency as $\rho(t) = \sum_w n_1^w / \sum_w \sum_q n_q^w$, where n_q^w is the total number of occurrences (tokens) of variant q for the word w at year t. (a) The orthography of German words that changed to "ss" ($q = 1$) from "ß" ($q = 2$) in the orthographic reform of 1996 (many words changed from "ss" to "ß" in the 1901 reform). (b) The transliteration of Russian names ending with the letter "в" when written in English (Latin alphabet), changed to an ending in "v" ($q = 1$) from endings in "ff" ($q = 2$) or "w" ($q = 3$) (e.g., $w =$ "Саратов" is nowadays almost unanimously written as "Saratov", but it used to be written also as "Saratoff" or "Saratow"). (c) The past form of the verb spill changed to its regular form "spilled" ($q = 1$) from the irregular form "spilt" ($q = 2$). The light curve shows the fit of Eq. (6.3). The estimated parameters a and b are (a) $\hat{a} = 0.218, \hat{b} = 0.000$ in 1901, and $\hat{a} = 0.229$, $\hat{b} = 0.000$ in 1996; (b) $\hat{a} = 0.000$, $\hat{b} = 0.099$; and (c) $\hat{a} = 0.001$, $\hat{b} = 0.030$. The corpus is the Google-ngram plotted in the minimum (yearly) resolution, see Appendix A.4 for details on the data and Sec. 6.2.2 for details on the fit.

to data of aggregated (macroscopic) S-curves because only very rarely one has access to detailed data at the individual (microscopic) level, see, e.g., Ref. [MZL12] for an exception.

Data alone is not enough to address the questions listed above, it is also essential to consider mechanistic models responsible for the change [Niy06, BBCM06, KGW08, BC12, PSD14]. Dynamical processes in language can also be described from the more general perspectives of evolutionary processes [BC12, Niy06, BR85] and complex systems [CFL09, BLT12, SCMF10]. In this framework, the adoption of new words can be seen as the adoption of innovations [Rog03, VA12, Bas69, Bas04, BCAKCC06, PSD14]. One of the most general and popular models of innovation adoption showing S-curves is the Bass model [Bas69, Bas04]. In its simplest case, it considers a homogeneous population and prescribes that the fraction of adopters (ρ) increases because those that have not adopted yet ($1 - \rho$) meet adopters (at a rate b) and are subject to an external force (at a rate a). The adoption is thus described by

$$\frac{d\rho(t)}{dt} = (a + b\rho(t))(1 - \rho(t)). \tag{6.2}$$

The solution (considering $\rho(t_0) = \rho_0$ and $\rho(\infty) = 1$) is

$$\rho(t) = \frac{a(1-\rho_0) - (a + b\rho_0)e^{(a+b)(t-t_0)}}{-b(1-\rho_0) - (a + b\rho_0)e^{(a+b)(t-t_0)}}. \tag{6.3}$$

It contains as limiting cases a *symmetric* S-curve (for $a = 0$) and an exponential relaxation (for $b = 0$). The fitting of Eq. (6.3) to the data in Fig. 6.3 leads to very different a and b in the three different examples, strongly suggesting that the S-curves are not universal and contain information on the adoption process. For instance, orthographic reforms are known to be exogenously driven (by language academies) in agreement with $b = 0$ obtained from the fit in panel (a).

In this section we investigate the shape and significance of S-curves in models of adoption of innovations and in data of language change. In particular, we estimate the contribution of endogenous and exogenous factors in S-curves, a popular question which has been addressed in complex systems more generally [SDGA04, CS08, AB04, MAAJ13]. The different values of a and b in Eq. (6.2) are an insufficient quantification, e.g., because they fail to indicate which factor is stronger. In Sec. 6.2.1 we introduce a definition for the relevance of different factors in a change based on the microscopic dynamics. We then show in Sec. 6.2.2 how this quantity can be exactly computed in different models and propose three different methods to estimate it from the time series of $\rho(t)$. Finally, we compare the accuracy and robustness of the methods using simulations of different network models in Sec. 6.2.3 and apply the methods to linguistic changes in Sec. 6.2.4.

6.2.1. Theoretical framework

Consider that $i = 1, \ldots, N \to \infty$ identical agents (**assumption 1**) adopt an innovation. The central quantity of interest for us here is $\rho(t) = N(t)/N$, the fraction of adopters at time t. We assume that $\rho(t)$ is monotonically increasing from $\rho_0 \equiv \rho(0) \approx 0$ to $\rho(\infty) = 1$ and agents after adopting the innovation do not change back to non-adopted status (**assumption 2**).

Endogenous and Exogenous Factors

In theories of language and cultural change, the importance of different factors is a topic of major relevance, e.g., Labov's internal and external factors [WLH68] and Boyd and Richerson's different types of biases in cultural transmission [BR85]. The first question we address is how to measure the contribution of different factors to the change. To the best of our knowledge, no general answer to this question has been proposed and computed in adoption models. As a representative case, we divide factors as endogenous and exogenous to the population. Mass media and decisions from language academies count as exogenous factors while grassroots spreading as an endogenous factor. In our simplified classification, Labov's internal (external) factors (to properties of the language [WLH68]) are counted by us as exogenous (endogenous), while Boyd and Richerson's [BR85] direct bias count as exogenous whereas the indirect bias and frequency-dependent bias count as endogenous.

Our proposal is to quantify the importance of a factor j as the number of agents that adopted the innovation because of j. More formally, let $g_i(t)$ be the adoption probability at time t for agent i (who is in the non-adopted status). We assume that g_i can be decomposed in contributions of the different factors j as $g_i(t) = \sum_j g_i^j(t)$, where $g_i^j(t)$ is the adoption probability of agent i at time t because of factor j. If t_i^* denotes the time agent i adopts the innovation, $g_i^j(t_i^*)/g_i(t_i^*)$ quantifies the contribution of factor j to the adoption of agent i (the adoption does not explicitly depends on $t < t^*$

and therefore values of $g_i^j(t)$ for $t < t^*$ are only relevant in the extent that they influence $g_i^j(t = t^*)$). In principle, the factor $g_i^j(t_i^*)/g_i(t_i^*)$ can be obtained empirically by asking recent adopters for their reasons for changing, e.g., for j=exogenous (endogenous) one could ask: *How much advertisement (peer pressure) affected your decision?*. We define the normalized quantification of the change in the whole population due to factor j as an average over all agents

$$G^j = \frac{1}{N}\sum_{i=1}^{N}\frac{g_i^j(t_i^*)}{g_i(t_i^*)}. \tag{6.4}$$

In order to show the significance of definition, Eq. (6.4), and how it can be applied in practice, we discuss how g_i^j and G^j can be considered in different models. Endogenous (endo) factors happen due to the interaction of an agent with other agents (internal to the population). They are therefore expected to become more relevant as the adoption progress (for increasing ρ). Exogenous factors (exo), on the other hand, are related to a source of information (external to the population) which has no dependence on ρ or time (**assumption 3**), see Fig. 6.4 for an illustration of the distinction between exogenous and endogenous spreading.

For simplicity, we report $G \equiv G^{\text{exo}}$ (since $G^{\text{endo}} = 1 - G^{\text{exo}}$).

Population dynamics models

Consider as a more general form of Eq. (6.2)

$$\dot{\rho}(t) \equiv \frac{d\rho(t)}{dt} = g(\rho(t))(1 - \rho(t)), \tag{6.5}$$

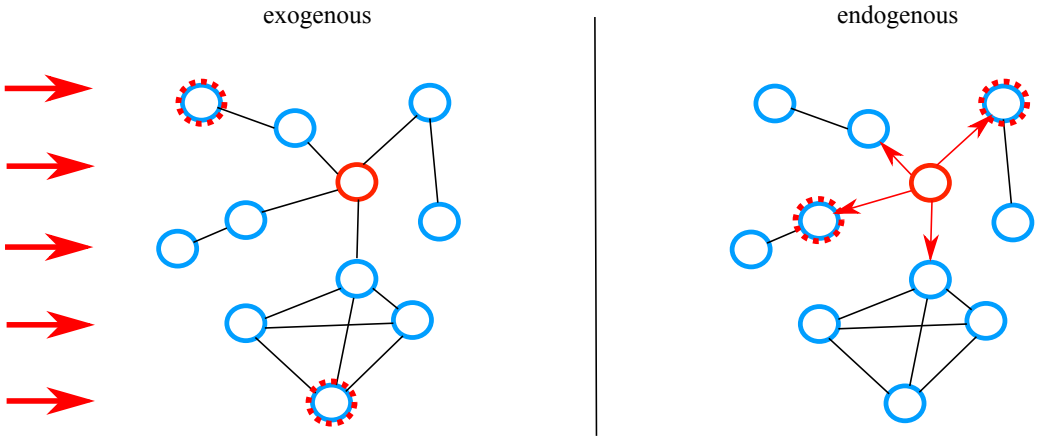

Figure 6.4.: Illustration of the distinction between exogenous and endogenous factors. Agents (nodes) are connected (links) in a given network. Starting from an agent that has adopted the innovation (red full circle) at a given time, in the next time step two non-adopters (blue full circle) become adopters (dotted red circle). (Left) The influence of exogenous factors (red arrows) does not depend on individual adopters, thus any node in the network can become infected with a given probability independent of the existing links. (Right) Endogenous factors (red arrows) are associated to spreading via links in the network, therefore, only nearest neighbours of the initially infected node can become infected in the next time step.

where $g(\rho(t))$ is the probability that the population of non-adopters $(1-\rho(t))$ switches from non-adopted status (0) to adopted status (1) at a given density of ρ. In epidemiology $g(\rho)$ is known as force of infection [HAF+10]. Since agents are identical (assumption 1) and $\rho(t)$ is invertible (assumption 2), we can associate $g_i^j(t_i^*)$ with $g^j(\rho)$ and $g_i(t_i^*)$ with $g(\rho)$. Introducing $g(\rho(t))$ from Eq. (6.5) in the continuous time extension of definition (6.4) we obtain:

$$G^j \equiv \int_0^1 \frac{g^j(\rho)}{g(\rho)} d\rho = \int_0^1 g^j(\rho) \frac{1-\rho}{\dot{\rho}} d\rho = \int_0^\infty \frac{g^j(t)}{g(t)} \dot{\rho}(t) dt. \tag{6.6}$$

This equation shows that the strength of factor j is obtained by averaging its normalized strength $g^j(\rho)/g(\rho)$ over the whole population or, equivalently, over time (considering the rate of adoption $\dot{\rho}(t)$).

When only exogenous and endogenous factors are taken into consideration, $g(\rho) = g^{\text{exo}} + g^{\text{endo}}$ in Eq. (6.5). Here, assumption 3 mentioned above corresponds to consider that the adoption happens much faster than the changes in the exogenous factors so that it can be considered independent of time. Therefore $g^{\text{exo}} = g(\rho = 0)$. Any change of g with ρ is an endogenous factor and $g^{\text{endo}}(\rho)$ increases with ρ because the pressure for adoption increases with the number of adopters.

For the case of the Bass model defined in Eq. (6.2), $g(\rho) = a + b\rho, g^{\text{endo}} = a, g^{\text{exo}} = b\rho$ and from Eq. (6.6) we obtain

$$G \equiv G^{\text{exo}} = \frac{a}{b} \log_e(\frac{a+b}{a}). \tag{6.7}$$

The correspondence of a and $b\rho$ to exogenous (innovators) and endogenous (imitators) is a basic ingredient of the Bass model [Bas69].[2] However, it is only through Eq. (6.7) that the importance of these factors to the change can be properly quantified. For instance, the case $a = b$ suggests equal contribution of the factors, but Eq. (6.7) leads to $G = \log_e 2 \approx 0.69 > 0.5$ and therefore shows that the exogenous factors dominate (are responsible for a larger number of adoptions than the endogenous factors). This new insight on the interpretation of the classical Bass model illustrates the significance of Eq. (6.4) and our general approach to quantify the contribution of factors.

Binary state models on networks

Another well-studied class of models inside our framework considers agents characterized by a binary variable $s = \{0, 1\}$ connected to each other through a network. We focus on models with a monotone dynamics (assumption 2), such as the Bass, Voter, and Susceptible Infected models, which are defined by the probability $F_{k,m}$ of switching from 0 to 1 given that the agent has k neighbours and m neighbours in state 1 [New10]. We use the framework of approximate master equations (AME) [Gle11, Gle13], which describes the stochastic binary dynamics in a random network with a given degree distribution P_k leading to the following system of ordinary differential equations for the fraction

[2] In our simple model, all agents are identical. The first adopters (innovators) are determined stochastically by the exogenous factor a, while agents adopting at the end of the S-curve (imitators) are more susceptible to the endogenous factor $b\rho$.

of susceptible $\{k, m\}$-nodes, $s_{k,m}$:

$$\frac{\mathrm{d}}{\mathrm{d}t} s_{k,m} = -F_{k,m} s_{k,m} - C\,(k-m)\, s_{k,m} + C\,(k-m+1)\, s_{k,m-1}, \tag{6.8}$$

where $m = 0, \ldots, k$ for each degree-class $k : P_k \neq 0$, $C = \langle (k-m)\, F_{k,m} s_{k,m} \rangle / \langle (k-m)\, s_{k,m} \rangle$, and $\langle \cdot \rangle = \sum_k P_k \sum_{m=0}^{k}$. From $s_{k,m}(t)$ we can calculate the timeseries for the total fraction of infected nodes, $\rho(t)$, according to:

$$\rho(t) = 1 - \sum_k P_k \sum_{m=0}^{k} s_{k,m}. \tag{6.9}$$

Assuming that at time t_0 a randomly chosen fraction of nodes, $\rho(t_0) = \rho_0$, is infected, we get as initial conditions for $s_{k,m}$ [Gle11]:

$$s_{k,m}(t_0) = \binom{k}{m} \rho_0^m \,(1-\rho_0)^{k-m}\,(1-\rho_0)\,. \tag{6.10}$$

In the Bass-model the probability of becoming infected is proportional to the number of neighbors that are already infected:

$$F_{k,m} = a + b\frac{m}{k}, \tag{6.11}$$

The one dimensional population dynamics model in Eq. (6.5) can be retrieved for simple networks (e.g., fully connected or fixed degree).

In a threshold-model a node becomes infected with probability 1 if the fraction of infected neighbors exceeds a certain threshold:

$$F_{k,m} = \begin{cases} a, & m/k > 1-b \\ 1, & m/k \leq 1-b \end{cases}, \tag{6.12}$$

The formulation of the spreading dynamics in the framework of AME allows us to calculate exactly the 'ground truth' of the exogenous and endogenous contributions for any given $F_{k,m}$. Following the approach above, we can now calculate exactly the individual contributions:

$$\begin{aligned} G^j &= \frac{1}{N} \sum_{i=1}^{N} \frac{g^j\,(t_i^*)}{g\,(t_i^*)}, \\ &= \frac{1}{N} \sum_k \sum_{i \in \{k\}} \frac{g^j\,(t_i^*, m_i^*, k)}{g\,(t_i^*, m_i^*, k)}, \\ &= \sum_k P_k \sum_{m=0}^{k} \int_0^\infty \frac{g^j\,(t, m, k)}{g\,(t, m, k)} \Delta_{k,m}\,(t)\,\mathrm{d}t, \end{aligned} \tag{6.13}$$

where $\Delta_{k,m}(t) = F_{k,m} s_{k,m}$ is the actual fraction of $\{k, m\}$-nodes that changed from susceptible to

infected at time t. Noting that the total rate of change is given by $g(k,m) = F_{k,m}$, it follows that

$$G^j = \sum_k P_k \sum_{m=0}^{k} \int_0^\infty g^j\left(t,m,k\right) s_{k,m}\left(t\right) \mathrm{d}t, \tag{6.14}$$

Assuming that the exogenous contribution is given by transitions that occur when no neighbor is infected, i.e. $g^{\mathrm{exo}}\left(k,m\right) = F_{k,0}$, the exogenous contribution yields:

$$G \equiv G^{\mathrm{exo}} = \sum_k P_k \sum_{m=0}^{k} \int_0^\infty F_{k,0} s_{k,m} \mathrm{d}t, \tag{6.15}$$

6.2.2. Time series estimators

In reality one usually has no access to information on individual agents and only the aggregated curve $\rho(t)$ is available. This means that G can not be estimated by Eqs. (6.4) or (6.15). Here we propose and critically discuss the accuracy and robustness of three different methods to estimate G from the S-curve $\rho(t)$ obtained from either empirical or surrogate data. All methods are inspired by the simple population model discussed above, but can be expected to hold also in more general cases.

We want to restrict ourselves to the analysis of the timeseries of the total fraction of adopters of the innovation, $\rho(t)$, with $\rho \in [0,1]$ given a set of N observations $D = \{t_i, \rho_i\, \sigma_i\}$ with $i = 1...N$, where t_i is the time, ρ_i the relative usage of one variant over the other, and σ_i the error associated to ρ_i (for details on the real data see Appendix A.4). Our starting point for all three methods is the assumption that the dynamics of the total fraction of adopters, $\rho(t)$, can be effectively described by a generalized population-dynamics model:

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}(t) = \left[1 - \tilde{\rho}\left(t\right)\right] g\left(\tilde{\rho}\left(t\right)\right), \tag{6.16}$$

which means that the rate of change of $\tilde{\rho}$ is determined by an arbitrary function $g(\tilde{\rho})$ only affecting the fraction of susceptibles, $1 - \tilde{\rho}$. Further, we want to account for the fact that the fraction of adopters is bounded by the two asymptotic values y_0 and y_1 such that $\rho(t \to -\infty) = y_0$ and $\rho(t \to \infty) = y_1$, which gives for the dynamics

$$\frac{\mathrm{d}}{\mathrm{d}t}\tilde{\rho}(t) = \begin{cases} \left[y_1 - \tilde{\rho}\right] g\left(\tilde{\rho}\right), & \tilde{\rho} \in [y_0, y_1] \\ 0, & \text{else} \end{cases}, \tag{6.17}$$

with an additional parameter t_0 setting a characteristic timescale, such that $\tilde{\rho}(t_0) = \frac{1}{2}(y_0 + y_1)$, which is equivalent to specifying the initial condition. Assuming a parametrization of $g(\tilde{\rho} \mid \theta)$ by the set of parameters θ we calculate the Least-Squared-Error, $\Delta(t_0, y_0, y_1, \theta)$ between data $D = \{t_i, \rho_i, \sigma_i\}$ and

the resulting curve $\tilde{\rho}(t \mid t_0, y_0, y_1, \theta)$ from our model

$$\Delta(t_0, y_0, y_1, \theta) = \sum_{i=1}^{N} \left(\frac{\rho_i - \tilde{\rho}(t_i \mid t_0, y_0, y_1, \theta)}{\sigma_i} \right)^2. \tag{6.18}$$

From this we can infer the most likely parameters $(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{\theta})$:

$$(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{\theta}) = \underset{(t_0, y_0, y_1, \theta)}{\operatorname{argmin}} \Delta(t_0, y_0, y_1, \theta) \tag{6.19}$$

Method 1: S- vs Exponential Curve (L)

We fit Eq. (6.3) by minimizing the Least-Square error with respect to the observed timeseries in the two limiting cases: (i) $a = 0$, symmetric S-curve (endogenous factors only) and (ii) $b = 0$, exponential curve (exogenous factors only). Assuming normally distributed errors (which generically vary in time) we calculate the likelihood of the data given each model [HTF09]. The normalized likelihood ratio L of the two models indicates which curve provides a better description of the data [BA02]. The critical assumption in this method (to be tested below) is to consider the value of L as an indication of the predominance of the corresponding factor, i.e $L > 0.5$ indicates stronger exogenous factors ($G > 0.5$) and $L < 0.5$ stronger endogenous factors ($G < 0.5$). This method does not allow for an estimation of G, but it provides an answer to the question of the most relevant factors. The two simple one-parameter curves are unlikely to precisely describe many real adoption curves $\rho(t)$. However, we expect that they will distinguish between cases showing a rather fast/abrupt start at t_0 (as in the exponential/exogenous case) from the ones showing a slow/smooth start (as in the S-curve/endogenous case). For this distinction, the $t \gtrapprox 0$ is the crucial part of the $\rho(t)$ curve because for $t \to \infty$ the symmetric S-curve approaches $\rho = 1$ also exponentially.

The two limiting $a = 0$ (endogenous) and $b = 0$ (exogenous) correspond to $g(\rho) = b(\rho - y_0)$ (endogenous) and $g(\rho) = a$ (exogenous), respectively, in which case we can solve Eq. (6.17) analytically which yields a four-parameter curve for each case:

$$\rho_{\text{exo}}(t \mid t_0, y_0, y_1, a) = \begin{cases} y_1 - \frac{1}{2}(y_1 - y_0)\, e^{-a(t-t_0)}, & t \geq t^* \\ y_0, & t < t^* \end{cases}, \tag{6.20}$$

$$\rho_{\text{endo}}(t \mid t_0, y_0, y_1, b) = y_0 + \frac{y_1 - y_0}{1 + e^{-b(y_1 - y_0)(t - t_0)}}, \tag{6.21}$$

with

$$t^* = t_0 - \frac{\ln 2}{a}. \tag{6.22}$$

Given our observational data D we can then find the best choice of parameters for each case and calculate the Least-Square-Error $\Delta_{\text{exo}}(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{a})$ and $\Delta_{\text{endo}}(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{b})$ according to Eqs. (6.18,6.19).

In order to decide which of the two models (endogenous or exogenous) fits the data better, we

employ the Bayesian information criterion (BIC) [Sch78] used in model selection [BA02, HTF09] which is given by

$$BIC_{\text{exo}} = \Delta(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{a}) + \log(N) K_{\text{exo}} \tag{6.23}$$

$$BIC_{\text{endo}} = \Delta(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{b}) + \log(N) K_{\text{endo}} \tag{6.24}$$

where $K_{\text{exo}} = K_{\text{endo}}$ is the number of fitted parameters in each model, and N the number of data points. From this we can calculate the relative likelihood, L_{exo} (L_{endo}), quantifying the evidence of the exogenous (endogenous) model among the selection of the two models (exogenous and endogenous) for the given data [BA02]:

$$L_{\text{exo}} = \frac{e^{-1/2 BIC_{\text{exo}}}}{e^{-1/2 BIC_{\text{exo}}} + e^{-1/2 BIC_{\text{endo}}}} \tag{6.25}$$

$$L_{\text{endo}} = \frac{e^{-1/2 BIC_{\text{endo}}}}{e^{-1/2 BIC_{\text{exo}}} + e^{-1/2 BIC_{\text{endo}}}} \tag{6.26}$$

with $L_{\text{exo}} + L_{\text{endo}} = 1$. In the last step we take the relative likelihood of the exogenous and the endogenous model as a proxy for their total influence in the spreading of the observed timeseries, i.e.

$$L \equiv G^{\text{exo}} = L_{\text{exo}} \tag{6.27}$$

with normalization $G^{\text{endo}} = 1 - G^{\text{exo}}$.

When analyzing surrogate data (see Sec. 6.2.3) the above problem becomes simpler since we know that $y_0 = 0$ and $y_1 = 1$ by construction. We further specify the initial condition for the spreading process, $\rho(t = t_0) = \rho_0$, which reduces the above curves to one-parameter models:

$$\rho_{\text{exo}}(t \mid a) = \begin{cases} 1 - (1 - \rho_0) e^{-a(t - t_0)}, & t \geq t^* \\ 0, & t < t^* \end{cases}, \tag{6.28}$$

$$\rho_{\text{endo}}(t \mid b) = \frac{1}{1 + e^{-b(t - t_0)}} \tag{6.29}$$

with $t_* = t_0 + \ln(1 - \rho_0)$.

Method 2: Mixed Curve ($\hat{G}$)

We fit Eq. (6.3) by minimizing the Least-Square error with respect to the timeseries and obtain the estimated parameters $\hat{a}$ and $\hat{b}$. By inserting these parameters in Eq. (6.7) we compute $\hat{G}$ as an estimation of G.

In this framework, we assume that, both, exogenous and endogenous driving is present in the spreading dynamics simultaneously, i.e. $g(\rho) = a + b\,(\rho - y_0)$, in which case Eq. (6.17) yields a

5-parameter curve for ρ:

$$\rho_{\text{mixed}}(t \mid t_0, y_0, y_1, a, b) = \begin{cases} \frac{-(a-by_0)(y_1-y_0)+y_1(2a+b(y_1-y_0))e^{[a+b(y_1-y_0)](t-t_0)}}{b(y_1-y_0)+(2a+b(y_1-y_0))e^{[a+b(y_1-y_0)](t-t_0)}}, & t \geq t^* \\ y_0, & t < t^* \end{cases}, \tag{6.30}$$

with

$$t^* = t_0 - \frac{\ln\left(2 + \frac{b}{a}\left(y_1 - y_0\right)\right)}{a + b\left(y_1 - y_0\right)}. \tag{6.31}$$

We note that the special case $a = 0$ yields $t^* \to -\infty$, which means that for all finite t: $\rho(t) > y_0$ and only in the limit $\rho(t \to -\infty) = y_0$. Given the data D we estimate the most likely parameters $(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{a}, \hat{b})$ using Eqs. (6.18,6.19).

For the given choice of $g(\rho) = a + (b - y_0)\rho$ we define the exogenous and the endogenous influence as

$$g^{\text{exo}}(\rho) = g(\rho = \hat{y}_0) = \hat{a}, \tag{6.32}$$

$$g^{\text{endo}}(\rho) = g(\rho) - g^{\text{exo}}(\rho) = \hat{b}(\rho - \hat{y}_0). \tag{6.33}$$

From this we can calculate the total exogenous and endogenous influence in the spreading process as the fraction of the population that switches at time t, $\dot{\rho}(t)$, weighted by the relative exogenous influence, $g^{\text{ext}}(\rho)/g(\rho)$, and relative endogenous influence, $g^{\text{endo}}(\rho)/g(\rho)$, respectively, integrated along the complete trajectory $\rho(t)$

$$\tilde{G} \equiv G^{\text{exo}} = \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{-\infty}^{\infty} \mathrm{d}t \dot{\rho}(t) \frac{g^{\text{exo}}(\rho(t))}{g(\rho(t))} \tag{6.34}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{\hat{y}_0}^{\hat{y}_1} \mathrm{d}\rho \frac{g^{\text{exo}}(\rho)}{g(\rho)} \tag{6.35}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \int_{\hat{y}_0}^{\hat{y}_1} \mathrm{d}\rho \frac{\hat{a}}{\hat{a} + \hat{b}\left(\rho - \hat{y}_0\right)} \tag{6.36}$$

$$= \frac{1}{\hat{y}_1 - \hat{y}_0} \frac{\hat{a}}{\hat{b}} \ln\left[\frac{\hat{a} + \hat{b}\left(\hat{y}_1 - \hat{y}_0\right)}{\hat{a}}\right] \tag{6.37}$$

and the normalization $G^{\text{endo}} = 1 - G^{\text{exo}}$.

When analyzing surrogate data (see Sec. 6.2.3) the above problem becomes simpler since we know that $y_0 = 0$ and $y_1 = 1$ by construction. For the surrogate data, see Sec. 6.2.3, we know that $y_0 = 0$ and $y_1 = 1$ by construction. We further specify the initial condition for the spreading process, $\rho(t = t_0) = \rho_0$, which reduces the above curve to a two-parameter model:

$$\rho_{\text{mixed}}(t \mid a, b) = \begin{cases} \frac{-a(1-\rho_0)+(a+b\rho_0)e^{(a+b)(t-t_0)}}{b(1-\rho_0)+(a+b\rho_0)e^{(a+b)(t-t_0)}}, & t \geq t^* \\ 0, & t < t^* \end{cases}, \tag{6.38}$$

with

$$t^* = t_0 + \frac{1}{a+b} \ln \frac{a\,(1-\rho_0)}{a+b\rho_0}. \tag{6.39}$$

Method 3: Nonparametric Curve ($\tilde{G}$)

We estimate $g(\rho)$ from Eq. (6.5) by calculating a (discrete) time derivative $\dot{\rho}$ at every point $\rho(t)$:

$$\hat{g}(\rho) := \frac{\dot{\rho}(t)}{y_1 - \rho(t)}. \tag{6.40}$$

From a (smoothed) curve of $g(\rho)$ we can infer the exogenous and the endogenous influence along the trajectory ρ:

$$g^{\text{exo}} = \hat{g}(\rho = y_0) \tag{6.41}$$

$$g^{\text{endo}} = \hat{g}(\rho) - \hat{g}(\rho = y_0) \tag{6.42}$$

and obtain an estimation $\tilde{G}$ of G from Eq. (6.6): which gives for the total exogenous and endogenous contribution

$$\hat{G} \equiv G^{\text{exo}} = \frac{1}{y_1 - y_0} \int \mathrm{d}\rho \frac{g^{\text{exo}}}{\hat{g}(\rho)} \tag{6.43}$$

with the normalization $G^{\text{endo}} = 1 - G^{\text{exo}}$. The advantage of this non-parametric method is that it is not a priory attached to a specific $g(\rho)$ and therefore it is expected to work whenever a population dynamics equation (6.5) provides a good approximation of the data.

When analyzing surrogate data (see Sec. 6.2.3) we can infer $g(\rho)$ *directly* with the timeseries $\rho(t)$ being sampled at a given resolution in discrete time, $t = (t_i)$ with $i = 1..N$, such that we can approximate the time derivate of $\rho(t)$ by finite differences, e.g.

$$\dot{\rho}(t_i) \approx \frac{\rho(t_{i+1}) - \rho(t_i)}{t_{i+1} - t_i} \tag{6.44}$$

for $i = 1...N-1$. Assuming that $\rho(t)$ is a monotone function in t, i.e. $t = t(\rho)$, we can express the time derivative as

$$\dot{\rho}(t) \overset{t=t(\rho)}{\longrightarrow} \dot{\rho}(\rho) \tag{6.45}$$

such that we can evaluate $\hat{g}(\rho)$, see Eq. (6.40), from the timeseries $\rho(t)$ and its derivative $\dot{\rho}$ via:

$$\hat{g}(\rho) := \frac{\dot{\rho}\,[t(\rho)]}{1-\rho} \tag{6.46}$$

However, for real data which is only available with a given resolution in t and is subject to fluctuations, the direct calculation of $\dot{\rho}$ in Eq. (6.40) does not lead to meaningful results. Instead, we want to

infer $g(\rho)$ *indirectly*, i.e. find a particular choice of $g(\rho)$ that yields the best description of the data by solving Eq. (6.17) for $\rho(t)$ and then applying Eqs. (6.18,6.19). Our approach is to parametrize $g(\rho)$ by means of a natural cubic spline $s(\rho)$ [HTF09]. Therefore, we divide the support of $g(\rho)$, $\rho \in [y_0, y_1]$, into n intervals of equal length $h = \frac{y_1 - y_0}{n}$, $\{[y_0 + (i-1)h, y_0 + ih\}$ for $i = 1...n$. In each interval i we define a cubic polynomial, such that the resulting curve $s_n(\rho)$ is piecewise-polynomial of order 4 and has continuous derivatives up to order 2. Furthermore, we restrict ourselves to natural cubic splines which implies that $s_n''(\rho = y_0) = s_n''(\rho = y_1) = 0$. The resulting spline $s_n(\rho)$ contains $n+1$ parameters $\theta = (\theta_i)$ with $i = 1...n+1$ and two additional parameters (y_0, y_1) specifying the asymptotic values for $\rho(t \to \pm\infty)$ which we denote by $s_n(\rho \mid (y_0, y_1, \theta))$. For any given n we can infer $\hat{g}_n(\rho) = s_n(\rho \mid \hat{y}_0, \hat{y}_1, \hat{\theta})$ by Eqs. (6.17,6.18,6.19), which requires an extra parameter t_0 setting a characteristic time scale of the change of $\rho(t)$ in time. In total, for a parametrization of $g(\rho)$ by a natural cubic spline on n intervals, we have $K = n+4$ parameters. Finally, the exogenous and the endogenous influence in the spreading are calculated via Eq. (6.43).The crucial step then is to decide which value to choose for n as we have to find a trade-off between the most accurate description of the data and the problem of overfitting known as model selection [HTF09]. We infer the best model by means of the Bayesian information criterion (BIC), which penalizes models with additional parameters according to:

$$BIC = \Delta + K \log N, \tag{6.47}$$

where Δ is the Least-Square error of the best fit of a given model according to Eqs. (6.18,6.19), K is the number of parameters estimated, and N is the number of datapoints. Due to computational constraints we restrict ourselves to the cases $n = 1, ...10$.

6.2.3. Application to network models

Here we investigate time series $\rho(t)$ obtained from simulations of models in which we have access to the microscopic dynamics of agents. Our goal is to measure G on different models and to test the estimators $(L, \tilde{G}, \hat{G})$ defined in the previous section.

Surrogate data

We consider two specific network models in the framework of AME described in Sec. 6.2.1, which are defined fixing the network topology (in our case random scale-free) and the function $F_{k,m}$ (the adoption rate of an agent having m out of k neighbours that already adopted) as [Gle13, New10]:

$$\text{Bass model: } F_{k,m} = a + b\frac{m}{k}, \tag{6.48}$$

$$\text{Threshold: } F_{k,m} = \begin{cases} a, & m/k < 1-b \\ 1, & m/k \geq 1-b \end{cases}. \tag{6.49}$$

In both cases, when no infected neighbor is present ($m = 0$), the rate is $F_{k,0} = a$ and therefore the parameter a controls the strength of exogenous factors. Analogously, b controls the increase of $F_{k,m}$ with m and therefore the strength of endogenous factors. Given a network and values of a and b, we obtain numerically both the timeseries $\rho(t)$ (using the AME formalism described in Sec. 6.2.1), and the strength of exogenous factors G from Eq. (6.15). Typically these models cannot be reduced to a one-dimensional population dynamics model and therefore the estimators $\hat{G}$ and $\tilde{G}$ (based on $\rho(t)$) differ from the actual G. As a test of our methods, we compare the exact G to L, $\hat{G}$ and $\tilde{G}$.

Numerical implementation

Given a degree-sequence $k \in [k_{\min}, k_{\max}]$, a degree distribution P_k, and one of the $F_{k,m}$ from Eqs. (6.11,6.12), we can solve the set of differential equations for $s_{k,m}$ numerically according to Eq. (6.8). We use scipy's [JOP+] odeint-implementation to get the timeseries $\rho(t)$ from Eq. (6.10) and the true exogenous and endogenous influence from Eq. (6.15) for a particular trajectory. We set as parameters $\rho_0 = \rho(t_0 = 0) = 10^{-3}$ and sample the trajectory $\rho(t)$ at discrete points $t \in \{t_0 + i \cdot dt\}$ for $i = 1..N$ with $dt = 0.01$ and $\rho(t = Ndt) \geq 1 - \rho_0$.

Results

In Fig. 6.5 we apply our time series analysis to the two models defined above with parameters $a = 0.1, b = 0.5$. Method 1 provides $L > 0.5$ in both cases, incorrectly identifying that the exogenous factor is stronger. Furthermore, $\tilde{G}$ (Method 3) provides a better estimation of G than $\hat{G}$ (Method 2). This is expected since the estimation $\hat{G}$ is based on a straight line estimation of $g(\rho)$, $(\hat{a} + \hat{b}\rho)$, while $\tilde{G}$ admits more general function, see Fig. 6.5, (b,d). The estimations are better for the Bass model than for the threshold dynamics, consistent with the better agreement between $\rho(t)$ and the fit of Eq. (6.3) in panel (a) than in panel (c).

In Fig. 6.6 we repeat the analysis of Fig. 6.5 varying the parameters a, b in Eqs. (6.48) and (6.49), while Eq. (6.15) gives the true value of G. The parameter space a, b is divided in two regions: one for which the exogenous factors dominate $G > 0.5$ (below the red dashed line $G = 0.5$) and one for which the endogenous factors dominate $G < 0.5$ (above the red dashed line $G = 0.5$). In the Bass dynamics the division between these regions corresponds to a smooth (roughly straight) line. In the threshold model a more intricate curve is obtained, with plateaus on rational values of b reflecting the discretization of the threshold dynamics in Eq. (6.49) (particularly strong for the large number of agents with few neighbors). A strong indication of the limitations of the L and $\hat{G}$ estimators is that the $L = 0.5$ (panel d) and $\hat{G} = 0.5$ (panel e) lines show non-monotonic growth in the a, b space. This artifact disappears using the $\tilde{G}$ estimator. Regarding the relative errors of the methods 2 and

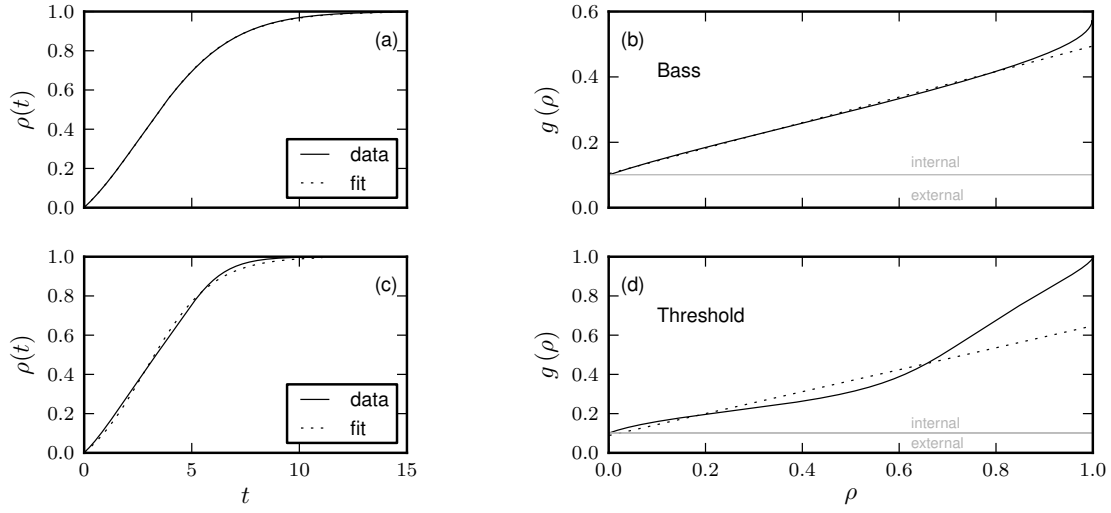

Figure 6.5.: Application of time series estimations to surrogate data. The Bass (a,b) and threshold (c,d) dynamics with parameters $a = 0.1$ and $b = 0.5$ were numerically solved in the AME framework for scale free networks (with degree distribution $P(k) \sim k^{-\gamma}$ with $\gamma \approx 2.47$ for $k \in [2, 50]$ such that $\langle k \rangle = 4$). (a,c) Adoption curve $\rho(t)$ (fraction of adopted agents over time). (b,d) Numerical estimate of $g(\rho)$, obtained from $\rho(t)$ by inverting Eq. (6.5). Dashed curves correspond to the fit of Eq. (6.3) to $\rho(t)$. Estimations of G correspond to the area between the horizontal gray line ($g(\rho) = \hat{a}$) and the solid ($\tilde{G}$) or dashed ($\hat{G}$) curves in (b,d). Results: Bass $G = 0.397, L = 0.999, \hat{G} = 0.415, \tilde{G} = 0.400$; Threshold $G = 0.347, L = 0.988, \hat{G} = 0.314, \tilde{G} = 0.352$.

3 (color code), the results confirm that $\tilde{G}$ is the best method and provides a surprisingly accurate estimation of G. Comparing the different models, the estimations for Bass are better than for the threshold dynamics (for the same parameters (a, b)). The minimum errors are obtained for $b \approx 0$ while for $a \approx 0$ maximum errors for both methods are observed.

When applying these methods to real data it is crucial not only to assess the accuracy of each method but also the robustness with respect to fluctuations and perturbations. For instance, method 3 requires the computation of the temporal derivative of ρ. In simulations this can be done exactly, however in empirical data, discretization (i.e. the time resolution of the available data) is unavoidable. Furthermore, fluctuations in the time series become magnified when discrete time differences are computed. In order to test these hypotheses, in Fig. 6.7 we test the robustness of Methods 2 and 3 against discretization in time – panels (a) and (b) – and population – panels (c) and (d) – for the same model systems. We observe that Method 3 is less robust than Method 2, showing a bias towards larger G for temporal discretization and broad fluctuations for population discretization. These findings can be expected to hold for other types of noise.

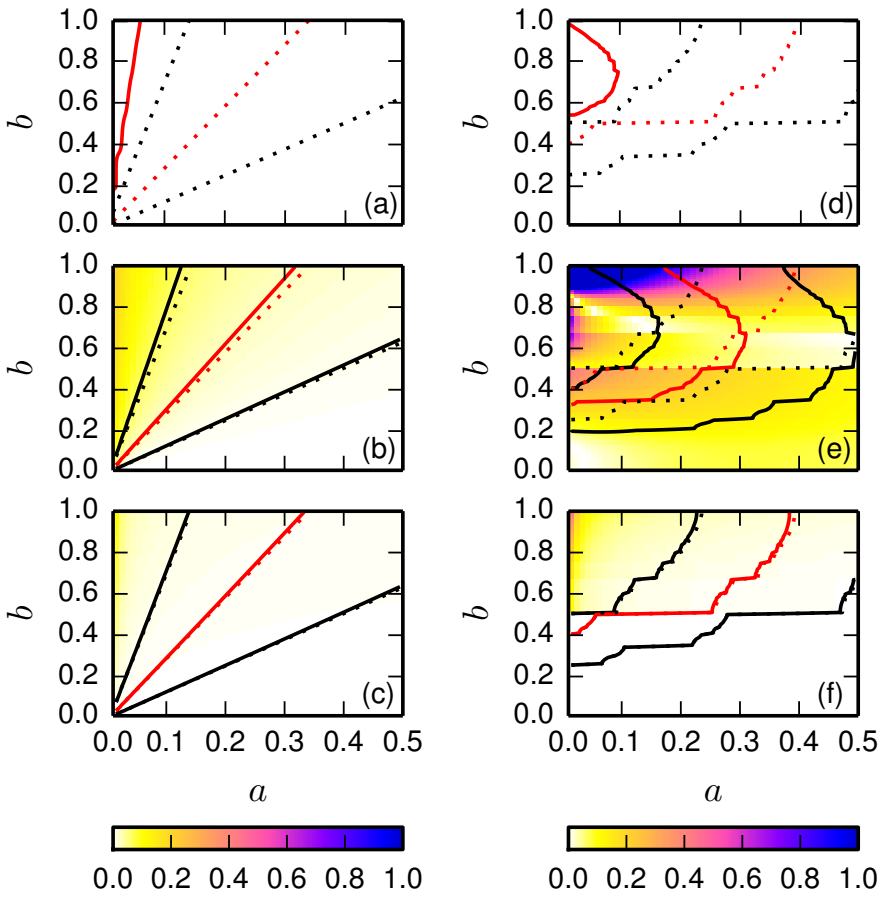

Figure 6.6.: Strength of endogenous factors G in the Bass [Eq. (6.48), panels a,b,c] and threshold [Eq. (6.49), panels d,e,f] models for different parameters a and b. The dashed lines correspond to values of a, b for which $G = 1/2$ (red), $G = 1/3$ (black below red), and $G = 2/3$ (black above red), computed from Eq. (6.15). The different panels show the estimations based on L (a,d), $\hat{G}$ (b,e), and $\tilde{G}$ (c,f). Solid lines indicate values of a, b for which values $1/2, 1/3$, and $2/3$ were obtained and should be compared to the corresponding dashed lines. The color code indicates the relative errors between the true value G and the estimated values $\hat{G}$ (b,e) and $\tilde{G}$ (c,f). The model dynamics was simulated for scale-free networks with the same parameters as in Fig. 6.5.

6.2.4. Application to data

We now turn to the analysis of empirical data taken from the Google-ngram corpus, see Appendix A.4. We focus on the three cases reported in Fig. 6.3.

Real data

a. German orthographic reforms: The 1996 orthography reform aimed to simplify the spelling of the German language based on phonetic unification. According to this reform, after a short vocal one should write "ss" instead of "ß", which predominated since the previous reform in 1901. This rule makes up over 90% of the words changed by the reform [Wik14b]. We combine all words affected by this rule to estimate the strength of adoption of the orthographic reform, i.e., $\rho(t)$ is the fraction of word tokens in the list of affected words written with "ss". Although following the reform was

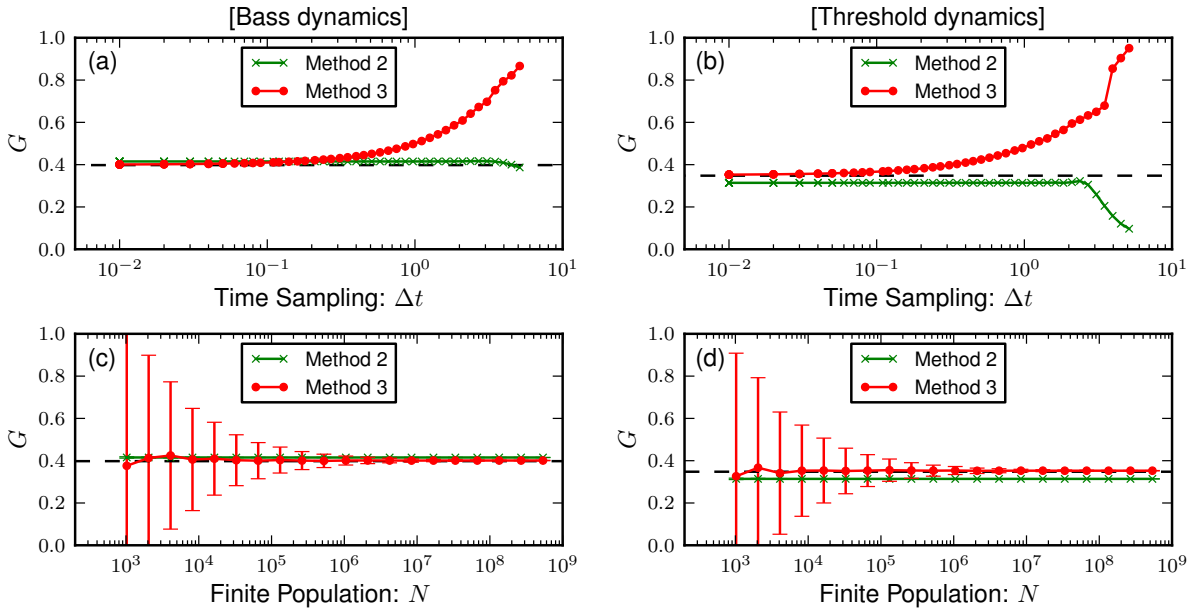

Figure 6.7.: Method 2 is more robust against perturbations than Method 3. Estimation of G in under-sampled versions of the timeseries used in Fig. (2) for Bass (left) and threshold (right) dynamics. The true G [Eq. (6.15)] is shown as a dashed line and Methods 2 and 3 are shown by symbols. (a,b) Under-sampling in time: achieved by varying the time resolution Δt of the timeseries, i.e., we sample $\rho(t)$ at times $\rho(t_0), \rho(t_0+\Delta t), \rho(t_0+2\Delta t), \ldots$. Resolution increases for $\Delta t \to 0$. (c,d) Under-sampling of the population N. The surrogate time series $\rho(t)$ in Fig. 6.5 assume $N \to \infty$. We consider time series for which only a finite population N is observed. The observed fraction of adopters is determined from N independent Bernoulli trials with probability $\rho(t)$. This corresponds to adding noise to each data point $\rho(t)$. Resolution increases for $N \to \infty$. For each N, we plot the average and standard deviation of G computed over 1,000 trials.

obligatory at schools, strong resistance against it led to debates even in the Federal Constitutional Court of Germany [Joh05]. For example, "six years after the reform, 77% of Germans consider the spelling reform not to be sensible [Wik14b]". These debates show that besides the exogenous pressure of language academies, endogenous factors can be important in this case also, either *for* or *against* the change.

b. Russian names: Since the 19th century there have been different systems for the romanization of Russian names, i.e. for mapping names from the Cyrillic to the Latin alphabet [Wik14c]. These systems can be seen as exogenous factors. Alternatively, imitation from other authors can be considered as endogenous factors. All of the systems suggest a unique mapping from letter "в" to "v" (e.g., Колмогоров to Kolmogorov). Variants to this official romanization system are "ff" or "w" (e.g., Kolmogorow and Kolmogoroff) which were used in different languages such as German and English. Here we study an ensemble of 50 Russian names ending in either "-ов" or "-ев" that were used often in English (en) and German (de). For each of these two languages, we combine all words (tokens) in order to obtain a single curve $\rho(t)$ measuring the adoption of the "v" convention.

c. Regularization verbs in English: A classical studied case of grammatical changes is regularization of English verbs [LMJ+07, Pin99]. From 177 irregular verbs in Old-English, 145 cases survived in

Middle English and only 98 are still alive [LMJ+07]. Irregular verbs coexist with their regular (past tense written by -ed) competitors, even if dictionaries may only present irregular forms [MSA+11]. Having an easier grammar rule or a rule aligned with a larger grammatical class are good motivations to use more often regular forms. Other potential exogenous factors which favor works against regularization can be dictionaries and grammars. However, there are also cases of verbs that become irregular [MSA+11, CPC+14]. We analyze 10 verbs that exhibit the largest relative change. In 8 cases regularization is observed.

Besides the linguistic and historical interest in these three cases, there are also two practical reasons for choosing these three simple spelling changes: (i) they provide data with high resolution and frequency; and (ii) they allow for an unambiguous identification of "competing variants", a difficult problem in language change [HCB+09]. The last point allows us to concentrate on the relative word frequency (as defined in the caption of Fig. 6.3) which we identify with the relative number of adopters $\rho(t)$ in the models of previous sections. The advantage of investigating relative frequencies, instead of the absolute frequency of usage of one specific variation, is that they are not affected by absolute changes in the usage of the word.

Numerical Implementation

The above mentioned methods require the minimization of the least-square error, see Eq. (6.19), in the space of parameters (t_0, y_0, y_1, θ). We find the most likely parameters $(\hat{t}_0, \hat{y}_0, \hat{y}_1, \hat{\theta})$ numerically using the 'L-BFGS-B'-algorithm [BLNZ95] from scipy's optimization package [JOP+]. The algorithm allows to impose additional constraints on a parameter x, such that we ensure that $x_{\min} \leq \hat{x} \leq x_{\max}$. In our case we choose the following constraints:

1. t_0 is unconstrained,
2. $0 \leq y_0, y_1 \leq 1$ since these parameters describe the asymptotic values of the fraction of adopters, i.e. $\rho(t \to \pm\infty)$,
3. $0 \leq a, b$ for method 1 and 2 considering positive exogenous and endogenous contributions,
4. $0 \leq \theta_i$ for $i = 1..n+1$ for method 3 in order to guarantee that $\hat{g}(\rho) \geq 0$.

Addressing the issue of local minima, for each timeseries we perform the minimization task 100 times with different randomly chosen initial conditions in parameter space and select the global minimum.

We calculate the confidence intervals from standard bootstrapping [HTF09], i.e. performing the same analysis for a number of B surrogate datasets obtained from random sampling with replacement of the original data (here $B = 200$).

Results

Fig. 6.8 shows estimations of the strength of exogenous factors G (using the methods of Sec. 6.2.2) in the three examples of linguistic change described above. In line with the definition proposed in

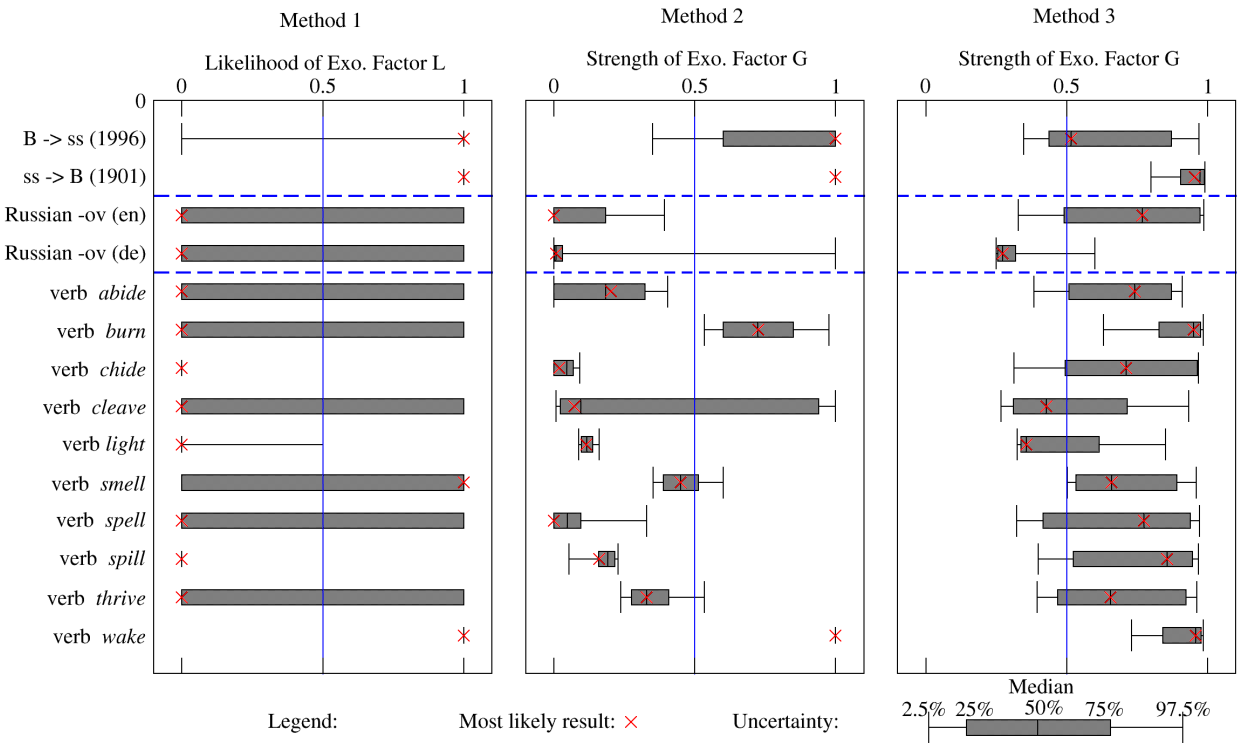

Figure 6.8.: Estimation of the strength of exogenous factors in empirical data. The red X indicates the estimated value obtained using the complete database. The box-plots (gray box and black bars) were computed using bootstrapping and quantify the uncertainty of the estimated value (from left to right, the horizontal bars in the boxplot indicate the 2.5%, 25%, 50%, 75%, and 97.5% percentile). Panels (a)-(c) show the estimations based on the three methods proposed in Sec. 6.2.2. (a) Method 1: the likelihood ratio L of the exponential fit (exogenous factors) in relation to the symmetric S-curve fit (endogenous factors). (b) Method 2: estimation $\hat{G}$ based on the fit of Eq. (6.3) and on Eq. (6.7). Method 3: estimation $\tilde{G}$ based on the general population dynamics model, Eq. (6.5).

Sec. 6.2.1, G is interpreted as the fraction of adoptions because of exogenous factors. Besides the most likely estimation obtained for the complete datasets (red X), we have performed a careful statistical analysis (based on bootstrapping) in order to determine the confidence of our estimations (gray box plots). We first discuss the performance of the three methods:

Method 1: The estimation of the likelihood L that the exponential fit (exogenous factors) is better than the symmetric S-curve fit (endogenous factors) resulted almost always in a categorical decision (i.e., $L = 0$ or $L = 1$). This is explained by the large amount of data that makes any small advantage for one of the fits to be statistically significant. Naively, one could interpret this as a clear selection of the best model. However, our bootstrap analysis shows that in most cases the decision is not robust against small fluctuations in the data (gray boxes fill the interval $L \in [0, 1]$). In these cases our conclusion is that the method is unable to determine the dominant factors (endogenous or exogenous).

Method 2: It generated the most tightly constrained estimates of G. The precision of the estimations of the strength of the exogenous factors G varied from case to case but remained typically much smaller than 1 (with the exception of the verb *cleave*). In all cases for which Method 1 provided a definite

result, Method 2 was consistent with it. This is not completely surprising considering that the fit of the curve used in method 2 has as limiting cases the curves used in the fit by Method 1. The advantage of Method 2 is that it works in additional cases (e.g., the Russian names), it provides an estimation of G (not only a decision whether $G > 0.5$), and it distinguishes cases in which both factors contribute equally (verb *smell*) from those that data is unable to decide (verb *cleave*).

Method 3: The results show large uncertainties and are shifted towards large values of G (in comparison to the two previous methods). In the few cases showing narrower uncertainties, an agreement with Method 2 is obtained in the estimated G (verbs *wake* and *burn*) or in the tendency $G < 0.5$ (Russian names in German). However, for most of the cases the uncertainty is too large to allow for any conclusion. The reason of this disappointing result is that Method 3 is very sensitive against fluctuations; compare the findings reported in Fig. 6.7, where we observed that Method 3 is less robust than Method 2, showing a bias towards larger G for temporal discretizations and broad fluctuations for population discretizations. These findings can be expected to hold for other types of noise and are consistent with our observations in the data.

We now interpret the results of Fig. 6.8 for our three examples (see Figs. 6.9, 6.10, 6.11, and 6.12 for the adoption curves of individual words):

a. Results for the **German orthographic reform** indicate a stronger presence of exogenous factors, consistent with the interpretation of the (exogenous) role of language academies in language change being dominant.

b. The **romanization of Russian names** indicates a prevalence of endogenous factors. Most systems that aim at making the romanization uniform have been implemented when the process of change was already taking place (The change starts around 1900 and first agreement is from 1950). Moreover, the implementation of these international agreements is expected to be less efficient than the legally binding decisions of language academies (such as in orthographic reforms).

c. The **regularization of English verbs** show a much richer behavior. Besides some unresolved cases (e.g., the verb *cleave*) the general tendency is for a predominance of endogenous factors (e.g., the verbs *spill* and *light*), with some exceptions (e.g., the verb *wake*).

In summary, in this paper we combined data analysis and simple models to quantitatively investigate S-curves of vocabulary replacement. Our data analysis shows that linguistic changes do not follow universal S-curves (e.g., some curves are better described by an exponential than by a symmetric S-curve and fittings of Eq. (6.3) leads to different values of $\hat{a}$ and $\hat{b}$). These conclusions are independent of theoretical models and should be taken into account in future quantitative investigations of language change.

In summary, non-universal features in S-curves suggest that information on the mechanism underlying the change can be obtained from these curves by considering simple mechanistic models of innovation adoption. Our results show a connection between the shape of the S-curves and the strength of different factors. Exogenous factors typically break symmetries of the microscopic dynamics and lead to asymmetric S-curves. Thus the crucial point in all methods is to quantify how abrupt (exogenous) or smooth (endogenous) the curve is at the beginning of the change. These findings and

the methods introduced in our approach – data analysis and measure of exogenous factors – can be directly applied also to other problems in which S-curves are observed [Rog03, VA12, Bas69, Bas04]. Since S-curves provide only a very coarse-grained description of the spreading of linguistic innovations in a population, the relevance of our results is to show that S-curves can be used to discriminate between different mechanistic models of the spreading mechanism and to quantify the importance of different factors known to act on language change. In view of the proliferation of competing models and factors, it is essential to compare them to empirical studies, which are often limited to aggregated data such as S-curves. Furthermore, quantitative descriptions of S-curves quantify the speed of change and predict future developments. These features are particularly important whenever one is interested in favoring convergence (e.g., the agreement on scientific terms can be crucial for scientific progress [KPW07] and dissemination [BGOB12]).

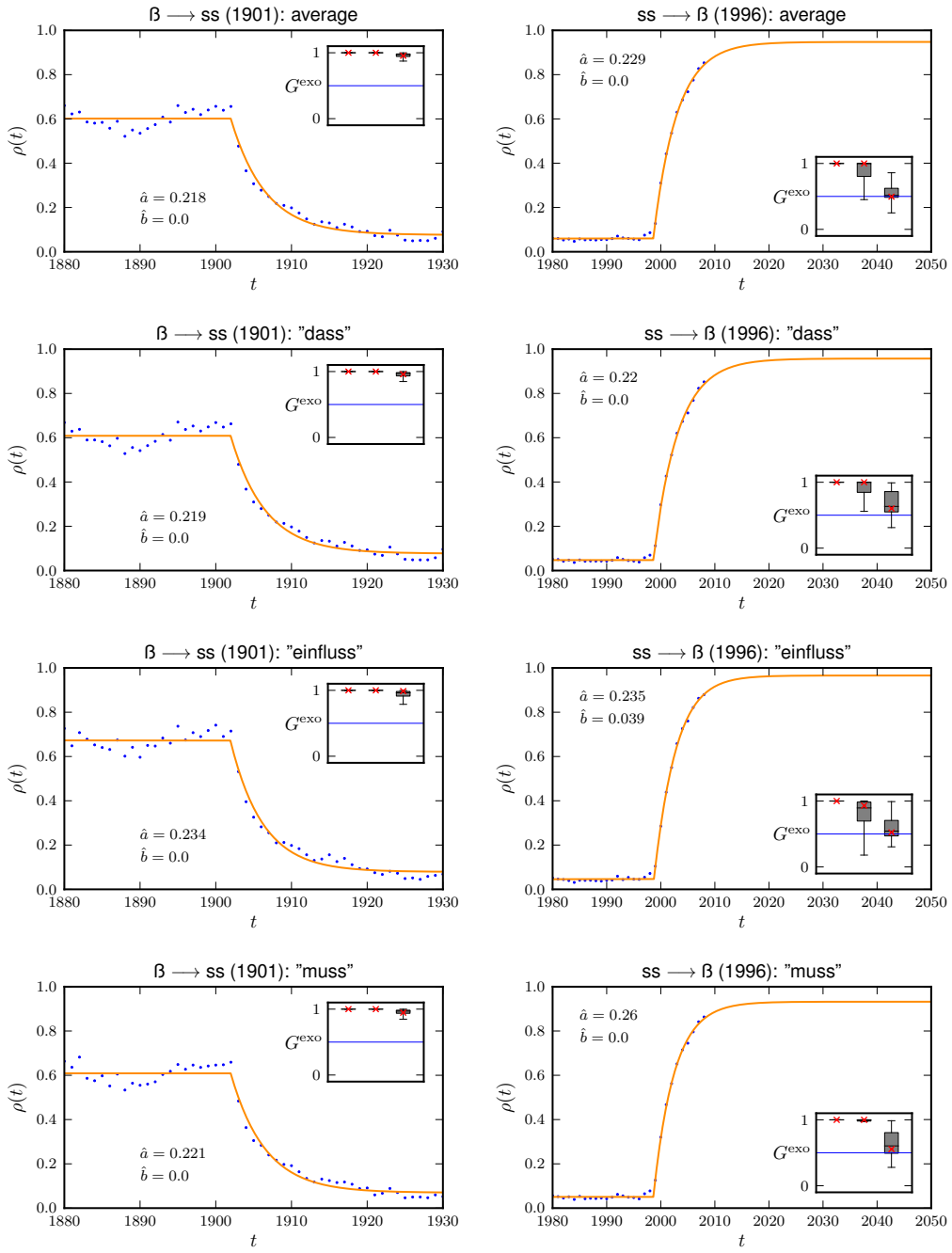

Figure 6.9.: Orthographic Reform of 1901 and 1996. The timeseries show the data (dots), the best fit of method 2 (line) with the values of its two parameters $\hat{a}$ and $\hat{b}$, and the boxplot for the estimation of G^{exo} (inset) for all three methods: method 1 (left), method 2 (middle), and method 3 (right) with the result for the full data (red cross) and the 97.5%-, 75%-, 50%-,25%-, and 2.5%-percentiles from bootstrapping (black lines).

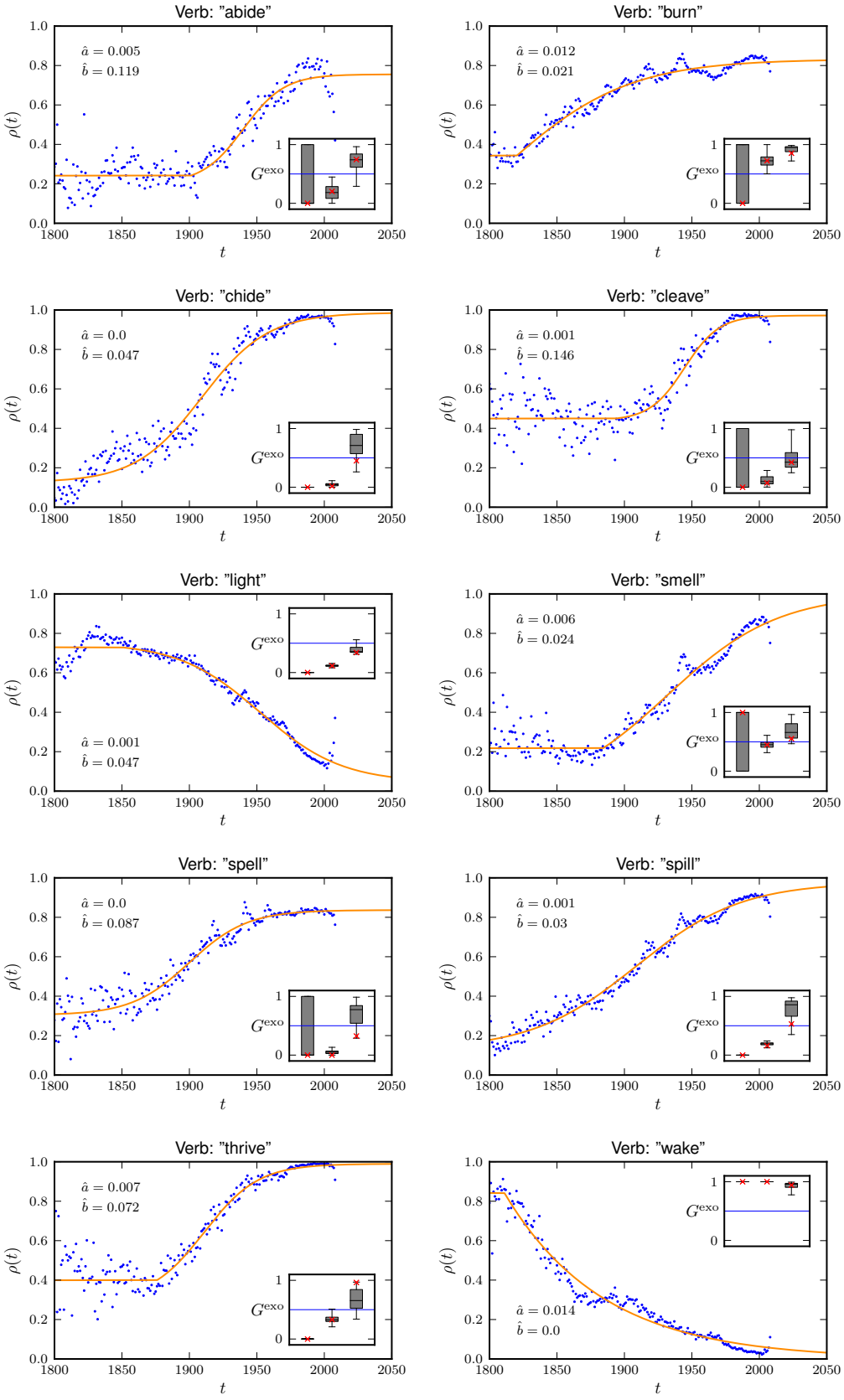

Figure 6.10.: Regularization of English Verbs. Description see Fig. 6.9.

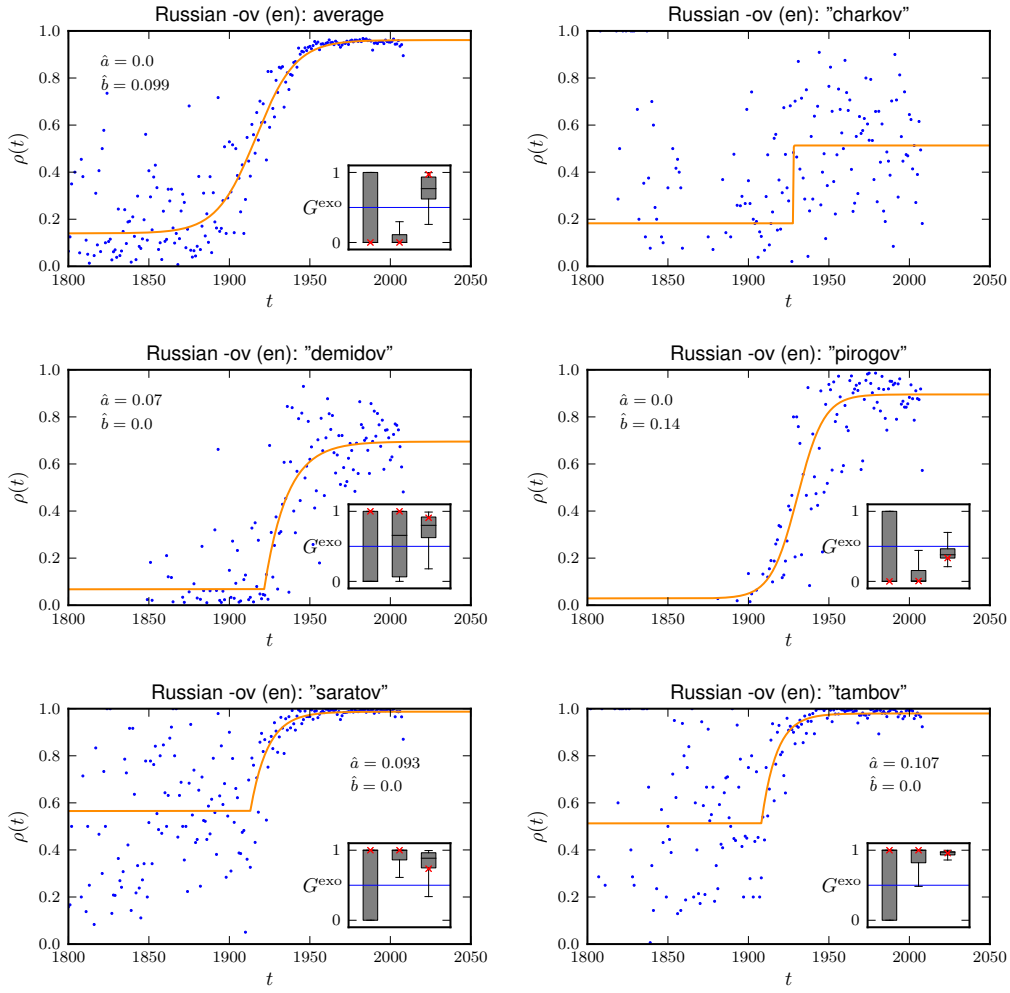

Figure 6.11.: Transcription of Russian -ov (en). Description see Fig. 6.9.

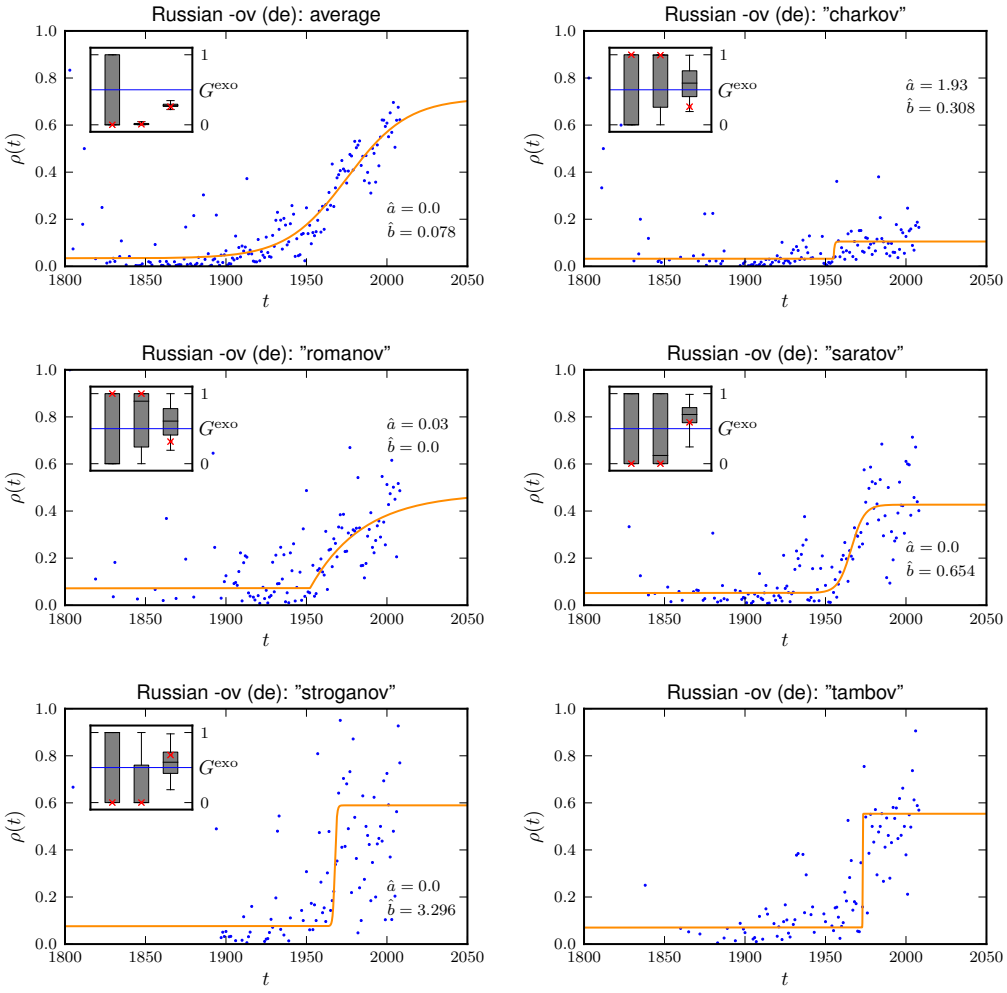

Figure 6.12.: Transcription of Russian -ov (de). Description see Fig. 6.9.

7. Conclusions

In this chapter we summarize and discuss the results presented in this thesis. In Sec. 7.1 we provide a detailed list of the novel results contained in this thesis. In Sec. 7.2 we discuss these results and present an outlook.

7.1. Summary and open problems

In the following we present a list summarizing the main novel results presented in this thesis as well as possible extensions and open problems that originate from this work.

List of specific results

1. Universal scaling in the word-frequency distribution (Secs. 3.1 and 3.2.2 and Ref. [GA13])
   - We found that the fat-tailed word-frequency distribution is best described by a two-parameter model with two power law regimes [Eq. (3.16)] where the values of the parameters are extremely robust with respect to time as well as the type and the size of the database under consideration depending only on the particular language. This constitutes the first rigorous statistical analysis on the double power-law generalization of Zipf's law confirming similar previous claims [FS01].
   - We proposed a simple growth model (Fig. 3.5) which allows for an interpretation of the two regimes in the word frequency distribution as a result of the existence of two classes of words: i) a finite number of core words, and ii) a virtually infinite number of noncore words.

2. Vocabulary growth (Secs. 3.2 and 4.2 and Refs. [GA13, GA14])
   - We proposed a simple stochastic process based on the Poisson usage of words establishing the connection between the word-frequency distribution and the vocabulary growth [Eq. 3.19], i.e. how the number of different words (V) depends on the total number of words (N). Assuming a double power law in the distribution of word frequencies, we found two distinct regimes in the sub-linear growth of the vocabulary [Eq. (3.32)] leading to a generalization of Heaps' law.
   - The prediction of our model showed remarkable agreement to the empirically measured vocabulary growth for different databases and over 9 orders of magnitude. From this we concluded that the growth in the vocabulary is driven mainly by the database size

and not by a change in vocabulary richness over time. This stands in contrast to recent studies claiming a steady increase of the vocabulary richness over time [MSA+11, Kle13], in which, in our opinion, the dependence on database size was not sufficiently addressed. Furthermore, we argued that the possible vocabulary can be regarded, for all practical purposes, to be infinite even though it is certainly bounded by combinatorial arguments due to a finite alphabet and word length.

- Analyzing the deviations from the average vocabulary growth, we found empirically that fluctuations are much larger than expected across different databases; in fact, relative fluctuations remain finite even in the limit $N \to \infty$. We formalized these observations in the framework of anomalous fluctuation scaling (Taylor's law), i.e. that the vocabulary is not self-averaging [Eq. (4.12)]. Extending the simple stochastic process discussed above by allowing for variability in the word frequencies, we showed analytically that the correlated usage of words (due to their semantic relation) can be responsible for this effect [Eq. (4.19) and Fig. 4.3].
- We applied these findings to propose a measure of vocabulary richness, which is unbiased with respect to the size of the texts [Fig.4.4].

3. Quantifying the similarity between texts (Secs. 4.3 and 6.1.2 and Ref. [GFCA15])
   - We quantified the similarity between two instances of text using the spectrum of divergences $\tilde{D}_\alpha$, which are a generalization of the Jensen-Shannon divergence based on the generalized entropies of order α [Eq. (4.33)]. We demonstrated that this approach is particularly useful in the comparison of texts that show fat-tailed distributions because it magnifies differences in the vocabulary at different scales of the frequency spectrum [Fig. 4.7].
   - Based on the notion of a generalized vocabulary [Eq. 4.41] we derived analytical expression for the bias and fluctuations in finite estimates of sample size N of the generalized entropy (H_α) and divergences ($\tilde{D}_\alpha$) typically decaying as N^β with $\beta < 1$ [Tab. 4.2]. We identified a critical value $\alpha^* \leq 2$ such that for $\alpha \geq \alpha^*$ the bias and the fluctuations decay as $1/N$ as expected for nonfat-tailed distributions, which motivates the pragmatic choice of $\alpha = 2$ in cases when the exponent of the fat-tailed distribution is unknown. Numerical simulations confirm the general validity of our results even when employing elaborated estimators of the corresponding entropies.
   - We applied the spectrum of divergences $\tilde{D}_\alpha$ to quantify how fast the vocabulary of a language is changing over time showing that $\tilde{D}_{\alpha=2}$ is best suited [Fig. 6.2]. As a result of the fat-tailed distribution of word frequencies, for $\alpha < 2$ (including the usual Jensen-Shannon divergence, $\alpha = 1$) the observed values of $\tilde{D}_\alpha$ are strongly influenced by finite-size effects (i.e. the bias) even though the size of the individual corpora is of the order of 10^9 word-tokens.

4. Topics models (Ch. 5 and Ref. [GPA15])

- We showed how the problem of topic modeling (finding large-scale structures in a collection of texts) can be mapped to the problem of community detection in complex networks.
- We compared state-of-the-art methods of topic modeling, i.e. Latent Dirichlet Allocation (LDA), and community detection, i.e. hierarchical stochastic block modeling (hSBM), showing that i) conceptually hSBM provides a much more general formulation of the problem solving many of the intrinsic limitations of LDA; and ii) hSBM provides better results in numerical experiments of artificial and real texts in terms of statistical model selection.

5. Spreading of linguistic innovations (Sec. 6.2 and Ref. [GGMA14])
   - We introduced a measure to quantify the strengths of different factors in the spreading of a linguistic innovation in a community of speakers [Eq. (6.4)]. Investigating microscopic models of spreading phenomena we find non-trivial relations between the parameters of the model and the resulting strength of individual factors [Fig. 6.6].
   - Discriminating between exogenous and endogenous influences to the population, we proposed a framework how to estimate these quantities from data available only on the macroscopic level (in the form of an average over the population).
   - Investigating three examples of competing linguistic variants, we showed that linguistic changes do not follow universal S-curves and that by applying the above framework one can obtain information on the underlying dynamical processes from macroscopic observables.

Open problems

- In the statistical analysis of power-law scaling relations (Ch. 3) it is desirable to include the effect of (long-range) correlations. This applies not only to language where these correlations are known to exist [ACE12], but also more generally, as recent works [CSN09, SP12] have increasingly questioned the validity of power laws and scaling relations in complex systems. However, especially if power-law scaling is viewed in the context of critical behaviour, successive observations are expected to be strongly correlated (i.e. not independent) which is not taken into account in standard approaches assessing the validity of these scalings (hypothesis testing).

- Although our analysis on the generalization of Zipf's law in the form of a double power law was shown to apply across different databases and languages (Sec. 3.1), the specific values for the parameters for each language (e.g. $\gamma = 1.77$ for the exponent in the second power law regime in English) still need to be explained.

- We studied the variability of word frequencies across different texts (Sec. 4). However, considering a single text as a succession of several smaller pieces of text it is possible to apply the same approach to investigate the topical variation within a given text, compare e.g. Ref. [MZ10]. This might provide a complementary view on approaches investigating texts as a time series of text constituents in order to understand their structural properties, e.g. burstiness [APM09] or long-range correlations [ACE12].

- In applying the hierarchical stochastic block model (hSBM) to the problem of finding large-scale structures in texts (Sec 5), the appearance of a phase transition between a detectable and an undetectable phase is still far from being understood, compare e.g. Ref [DKMZ11a, ZMN15] for recent approaches in simpler formulations of the stochastic block model. Additional insight might come from searching for analogies to the phase transition observed in spreading dynamics (e.g. the Susceptible-Infected-Susceptible model) on networks [PSCvV15], i.e. the existence of two regimes in which a disease either affects the whole network or dies out (the transition being called the epidemic threshold). The dependence of the epidemic threshold on the topology of the network has been well-understood in recent works [BnCPS13], especially the disappearance of the threshold if the network exhibits a scale-free degree distribution.

- In the problem of spreading of linguistic innovations in a community of speakers (Sec. 6.2) our approach is insofar limited as it only considers binary-state dynamics (an adopter either uses an innovation or not) on random networks. It would, therefore, be desirable to incorporate more realistic aspects into the microscopic description of the spreading phenomena, especially if one tries to compare its predictions with empirical data. For example, this could be achieved by i) allowing speakers to choose between different options with a given probability, e.g. as formulated in the utterance selection model [BBCM06]; or ii) including the notion of communities or clustering in the topology of the network, which is a crucial aspect of many empirical social networks.

7.2. Discussion and outlook

In this thesis we investigated the statistical and dynamical processes underlying the co-existence of universality and variation in word statistics. Our approach was guided by bridging the gap between empirical analysis and theoretical modeling combining careful statistical analysis of large records of written text with analytical and numerical studies of stochastic models in the form of generative processes. Validating that the distribution of word frequencies shows a remarkable degree of universal structure, we explored the consequences of the presence of fat-tailed distributions in the estimation of the size of the vocabulary and the quantification of the similarity between different texts across topics and time. Going beyond the appealing idea of universality, we showed that it is necessary to consider the variability of word frequencies in order to account for the fluctuations observed in empirical data and in order to describe variation of language across topics or language change over time.

In Sec. 3.1 we showed that the distribution of word frequencies is best described (among a set of generic descriptive models) by a double power law (i.e. two regimes with different power-law scalings), in which the fitted parameters of the model are extremely robust with respect to the type, size, or time of the database and only depend on the language. Taken as a signature of the universal structure in natural language, this finding is remarkable in the view of many recent works questioning the empirical support for the ubiquitous claims of power-law scalings in complex systems [CSN09, SP12].

Although our model is not valid in the strict sense of conventional hypothesis testing (which ignores the role of (long-range) correlations known to be present in texts [ACE12]), it provides the best simple description capturing the main statistical features of the data. We believe this improved interpretation of universal scaling and the statistical methodology we employed are applicable also in cases beyond language.

Besides the generalized Zipf's law describing the word-frequency distribution, we observed two additional universal scaling laws in the context of the vocabulary growth, i.e. Heaps' and Taylor's law. We obtained a unifying perspective on the simultaneous appearance of these scaling laws from a stochastic process formulated as a sampling problem. In this, we showed that it is necessary to consider the variability of word frequencies due to the topical aspects of language (i.e. texts coming from different authors, disciplines, or times). More generally, our probabilistic framework allowed for a calculation of the expected fluctuations, which is crucial for a meaningful interpretation of these universal laws. The picture that emerges from this approach is that, on average, these universal laws are extremely robust; however, fluctuations around these laws are typically much larger than expected from simplifying assumptions (e.g., independence or lack of correlations). This finding indicates that the constraints imposed by the universal structure are not as tight as one could expect. Beyond the cases considered here, these results provide a theoretical framework for studying fluctuations in the growth of the number of unique items investigated also in a range of other complex systems, e.g. ecology [Bra82], collaborative tagging [CBB$^+$09], networks [KR13], or in the general context of innovation dynamics [TLSS14].

Besides its consequences on the expected fluctuations, we analyzed the variability of word frequencies in the context of language change and topic models.

On the level of individual words, our analysis of the temporal spreading of linguistic innovations indicates that the adoption pattern do not follow universal S-curves. Starting from empirical time series describing the macroscopic behaviour (i.e. the total fraction of speakers that use the new variant), we showed how to obtain information on the underlying dynamics from the analysis of generic microscopic models of spreading phenomena. This approach goes beyond traditional analysis in which simple microscopic models are analyzed in order to understand the main qualitative features observed in spreading phenomena, e.g. [BC12]. It is also in line with recent approaches modeling the spreading of adoptions in general, e.g. Ref. [WPCG$^+$14] compared quantitatively the predictions of different spreading models with empirically observed S-curves of the number of adopters of a technological innovation in order to infer the best underlying microscopic models in terms of model selection (instead of assuming its validity in the first place).

On a larger scale, we addressed the question of how fast the vocabulary of a language changes over time by comparing the difference of word frequencies between two instances of texts from different times. For this, we proposed information-theoretic similarity measures based on the generalization of the Shannon entropy (the entropy of order α) yielding a spectrum of divergences. The use of this spectrum was motivated by the universal (fat-tailed) distribution of word frequencies, the generalized Zipf's law discussed above. On the one hand, we showed how different α's magnify different scales

in the frequency spectrum yielding additional information on the similarity between two texts. On the other hand, we calculated analytically that the convergence of systematic and statistical errors is often much slower than $1/N$ (where N is the text length). This illustrated the difficulty in obtaining accurate estimates due to the large number of low-frequency symbols even for very large databases ($N \gtrsim 10^9$ word-tokens). These results do not only have direct impact on other applications coming from computer science, e.g. authorship attribution or document classification, but can be viewed in the much more general context of quantifying the similarity of symbolic sequences. In this view, the significance of our results stem from the fact that many problems from other complex systems involving symbolic sequences show fat-tailed distributions in the frequencies of symbols, e.g., DNA [MBG+94], gene expression [FK03], or music [SCBn+12]).

Considering the variability of word frequencies not over time but across texts from different authors or disciplines we investigated the idea of topic models aiming to find large-scale structures in a collection of texts. In this we combined topic models coming from machine learning with the idea of community detection in complex networks showing and exploiting the analogies between the two approaches from both fields. On the one hand, we were able to propose a much more general formulation of the problem using stochastic block models. These models have recently become the subject of investigation in the framework of statistical physics [KN11, DKMZ11b], leading to a better understanding of the underlying mechanism of the proposed algorithms. On the other hand, we emphasized that the universal scaling laws in natural languages lead to new challenges in the problem of community detection in complex networks. For example, approaches in community detection often assume that the networks at hand are sparse, i.e. the number of edges scales linearly with the number of nodes as the size of the network is increased. However, this is not the case for the bipartite network composed of words and documents due to the sub-linear growth of the vocabulary (Heaps' law).

Appendix

A. Databases

A.1. Google-ngram

The data obtained from the Google-ngram database [MSA+11] contains the number of occurrences of words used in millions of printed books in the period 1500-2008 with a yearly resolution. We filter the data in two steps. First, we decapitalize each word (e.g. 'the' and 'The' are counted as the same word) and further restrict ourselves to words consisting uniquely of letters present in the alphabet of the corresponding language and the symbol " ' " (apostrophe). This is meant as a conservative approach in order to minimize the influence of foreign words, numbers (e.g. prices), or scanning problems which are present in the raw data. In the second step, when constructing yearly data $y(t)$, i.e., words present in books published in year t, we include only those words in the database $y(t)$, which appear more than 40 times in that particular year. In the same way, for the cumulative data $Y(t)$ we include only those words, which appeared more than 40 times until time t. In this way we avoid a possible bias due to the filtering applied in the construction of the raw data (words had to appear more than 40 times in all times in order to be included in the database [MSA+11]). As an example of possible bias, in case we had not applied this filter, take two words (called '1' and '2') with $n_1(t) = n_2(t) = 21$ occurrences in year t. If now $\forall t' \neq t : n_1(t') = 0$ and $\exists t'' \neq t : n_2(t'') > 20$, word '2' would be present in the raw data whereas word '1' would be not. As a result we would only include word '2' in the yearly database $y(t)$. With our additional filter neither word '1' nor word '2' appears in the yearly database $y(t)$.

In Fig. A.1 we show the resulting database size for the yearly data $y(t)$ and the cumulative data $Y(t) = \sum_{t'=to}^{t} y(t')$ in terms of word-tokens and word-types for English, French, Spanish, German, and Russian. In this context word-type refers to the number of distinct words, whereas word-token refers to the total number of words.

For the yearly database $y(t)$ we use data in the period $t \in [1805, 2000]$, because as already indicated in [MSA+11], the database composition may have changed in a noncontinuous way at $t \approx 1800$. This claim is supported in Fig. A.2, where we calculate Kendall's rank correlation coefficient $\tau[y(t), y(t')]$ between the common types of the database $y(t)$ and $y(t')$ for $1500 \leq t \leq t' \leq 2000$ as

$$\tau[y(t), y(t')] = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}, \tag{A.1}$$

where n is the total number of common elements, n_c the number of concordant, and n_d the number

of discordant pairs between the two databases with respect to the ranking of frequencies. Clearly, at $t = 1800$ a noncontinuous change in τ can be identified, from which we conclude that database composition is dramatically different in the years before and after $t = 1800$. In order not to be affected by this change the yearly data $y(t)$ is only considered in the period $t \in [1805, 2000]$. However, in order to take advantage of the full size of the database, the cumulative data $Y(t)$ is constructed taking into account all the years prior to $t = 1805$.

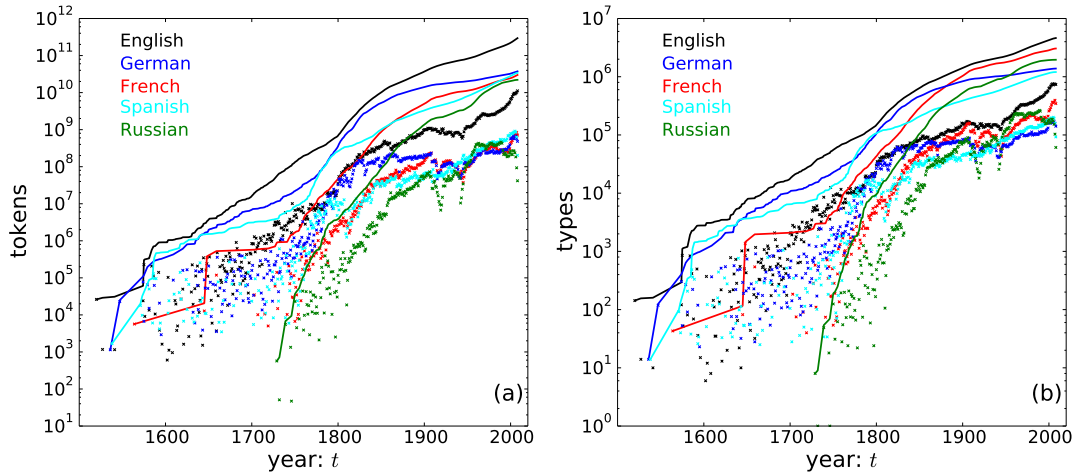

Figure A.1.: Size of the database after filtering. a) Number of tokens for yearly data $y(t)$ (x-symbols) and cumulative data $Y(t)$ (line). b) Number of types for yearly data $y(t)$ (x-symbols) and cumulative data $Y(t)$ (line). Each language is marked by a different color.

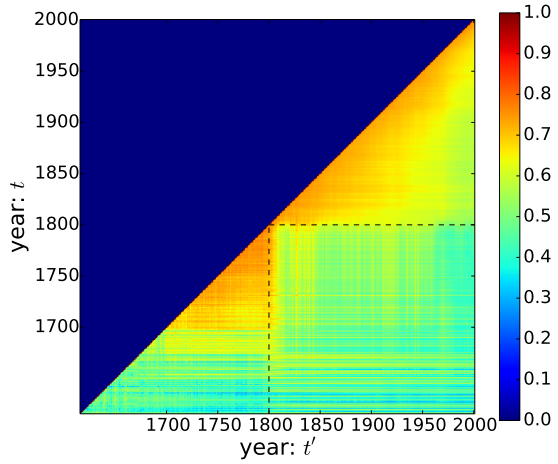

Figure A.2.: Correlation between data in different years for English. Kendall's rank correlation Eq. (A.1) between yearly data $y(t)$, $y(t')$ for $t, t' \in [1500, 2000]$ with $t \leq t'$. The dashed lines show $t = 1800$ and $t' = 1800$ where a noncontinuous change in the correlation occurs.

A.2. Wikipedia

The Wikipedia dataset contains a complete snapshot of the English Wikipedia [Wik] consisting of the full text of each individual article. We filter the Wikipedia database in three steps. First, using the WikiExtractor developed by the University of Pisa Multimedia Lab [oPML], we store only the plain text, neglecting any additional information or annotation such as images, tables, tags, references, or lists. In a second step we remove all punctuation characters (e.g. ",", ";", or "{") and cut the text into words at the whitespace characters in a similar manner as described in the construction of the Google-ngram database [MSA+11]. The final step consists of decapitalizing each word and restricting ourselves only to words consisting uniquely of the letters $a-z$, a filter we also applied to the Google-ngram database (see Sec. A.1). The resulting database comprises $3,743,306$ articles and consists of $\approx 3.7\ 10^6$ types and $\approx 1.3\ 10^9$ tokens.

A.3. PlosOne

The PlosOne database consists of all $76,723$ scientific articles published in the journal PlosOne which were accessible via the API at the time of the data collection [API] from which we extract the full text. We filter the data by decapitalizing each word and restricting ourselves only to words consisting uniquely of the letters $a-z$, a filter we also applied to the Google-ngram database (see Sec. A.1).

A.4. Time series of language change

The data for the timeseries is obtained from the latest version of the Google-ngram database [LMA+12], which is an extension of the original data [MSA+11] enriched with more data and syntactic annotations. Given two linguistic variants denoted by '1' and '2', we count the total number of occurrences of each variant $n_1(t)$ and $n_2(t)$ in each year $t \in [1800, 2008]$ irrespective of its capitalization. From this we can calculate $\rho(t)$, the relative usage of variant '1' over variant '2':

$$\rho(t) = \frac{n_1(t)}{n_1(t) + n_2(t)}. \tag{A.2}$$

We associate an error $\sigma_\rho(t)$ to each data point $(t, \rho(t))$, which we split into two parts, i.e.

$$\sigma_\rho^2 = \sigma_s^2 + \sigma_0^2 \tag{A.3}$$

in which σ_s is due to finite sampling of the data, and σ_0 subsumes additional uncertainties from potential exogenous perturbations which are not due to finite sampling effects. The introduction of the latter is necessary, because only considering the finite-sampling effect does not account for the observed fluctuations in the frequency of the most common words, which we assume to be stationary.

The effect of finite sampling, σ_s, is approximated by assuming that n_1 and n_2 are the outcomes of a binomial process with $n = n_1 + n_2$ samples where variant '1' is drawn with probability $p = n_1/N$

and variant '2' is drawn with probability $1-p$. From this we can calculate the error σ_s:

$$\sigma_s(t)^2 = \frac{n_1(t)n_2(t)}{(n_1(t)+n_2(t))^3}. \tag{A.4}$$

For the estimation of σ_0, which we treat as constant and independent of the sample size $n(t) = n_1(t)+n_2(t)$ in each year, we look at the timeseries of the relative frequency of the most frequent word, "the", in the English language. Assuming that this timeseries is stationary, we estimate σ_0, such that 95% of the points of the timeseries lie within the 95% confidence interval assuming Gaussian errors according to Eq. (A.3). This gives $\sigma_0^2 = 0.002$, which we take as a lower bound for unknown, exogenously driven fluctuations which have to be considered in order to account for observed fluctuations in the timeseries beyond finite-sampling effects.

Orthographic reform

In this section we focus exclusively on the competition between the letters 'ß' (s-sharp) and 'ss' encoding the sound for voiceless s in the German orthography. The official set of rules concerning the usage of each variant was changed twice in the orthographic reforms of 1901 and 1996 [Joh05]. We investigate the usage of each variant over time for 2960 words as being affected by the orthographic reform of 1996 [Can]. For each of these words we count in each year the number of times it occurred with variant 'ß', $n_{\text{ß}}(i,t)$, and the number of times it occurred with variant 'ss', $n_{\text{ss}}(i,t)$. We consider the timeseries of four representative cases: (i) the average over all words, $\rho_{\text{avg}}(t)$; (ii) the most frequent word 'dass', $\rho_{\text{dass}}(t)$; (iii) the most frequent verb 'muss', $\rho_{\text{muss}}(t)$; and (iv) the most frequent noun 'einfluss', $\rho_{\text{einfluss}}(t)$. We calculate the average relative usage of one variant as an average over all tokens, i.e.

$$\rho_{\text{avg}}(t) = \frac{N_{\text{ss}}(t)}{N_{ss}(t)+N_{\text{ß}}(t)} \tag{A.5}$$

with $N_{\text{ss}}(t) = \sum_{i=1}^{2960} n_{\text{ss}}(i,t)$ and $N_{\text{ß}}(t) = \sum_{i=1}^{2960} n_{\text{ß}}(i,t)$. The respective relative usage for the words 'dass', 'muss', 'einfluss' is calculated as follows:

$$\rho_{\text{word}=i}(t) = \frac{n_{\text{ss}}(i,t)}{n_{\text{ss}}(i,t)+n_{\text{ß}}(i,t)}. \tag{A.6}$$

The frequency of a word, $f_{\text{word}}(t)$, is measured as the relative weight of both variants combined:

$$f_{\text{word}=i}(t) = \frac{n_{\text{ß}}(i,t)+n_{\text{ss}}(i,t)}{\sum_j n_{\text{ß}}(j,t)+n_{\text{ss}}(j,t)} \tag{A.7}$$

Russian names

In this section we focus on two Russian surname-suffixes: 'ов' and 'ев'. The letter 'в' has been written in Roman script languages like English and German by 'v', 'w' or 'ff'. Here we consider the

competition between the letter 'v' and two others together 'w'+'ff'; while all the official standard systems suggest 'v' for 'в'. We investigate the usage of each variant over time for 50 common Russian surnames which are listed below.

German: Charkov, Saratov, Romanov, Stroganov, Tambov, Pirogov, Godunov, Katkov, Aksakov, Demidov, Semenov, Lermontov, Saltykov, Kornilov, Stepanov, Lobanov, Bulgakov, Krylov, Melnikov, Annenkov, Turgenev, Kostomarov, Filatov, Grekov, Putilov, Titov, Vinogradov, Danilov, Sobolev, Nikiforov, Kamenev, Novikov, Kondakov, Martynov, Rykov, Melikov, Platonov, Karpov, Lazarev, Balabanov, Krasnov, Nabokov, Dolgorukov, Kirov, Leonov, Maklakov, Naumov, Frolov, Mitrofanov, Fedotov

English: Saratov, Demidov, Pirogov, Tambov, Charkov, Katkov, Kornilov, Lazarev, Novikov, Melikov, Lermontov, Aksakov, Godunov, Turgenev, Menshikov, Stepanov, Vinogradov, Semenov, Kutuzov, Lebedev, Suvorov, Lomonosov, Mendeleev, Lavrov, Melnikov, Lobanov, Annenkov, Volkhov, Balakirev, Lvov, Bazarov, Shuvalov, Grigoriev, Titov, Yakov, Nekrasov, Mikhailov, Gorchakov, Morozov, Zubov, Chekhov, Sakharov, Dragomirov, Andreyev, Danilov, Chirikov, Yermolov, Bulgakov, Vasiliev, Saltykov

To make this lists, the primary list of common Russian surnames ending 'ов' and 'ев' was created according to the English Wikipedia pages including: List of surnames in Russia, List of Russian-language writers, scientists, composers, leaders of the Soviet Union and Marshal of the Soviet Union; Also list of cities and towns in Russia was counted in this list. The words which have been used at least a) one time between 1800 and 2008 and b) 10 times for more than 100 years (75 years for German data) in this period were included in the initial list. Then in order to guarantee that these words are right competitors we applied the following filtering:

- The words whose first letter were written mostly by small letters instead of capital letters ($\frac{\sum_{t=1800}^{2008} f_{\text{word}}^{\text{small}}(t)}{\sum_{t=1800}^{2008} f_{\text{word}}^{\text{captal}}(t)} \geq 0.01$) were excluded.
- The names like *Gorbachev* which has a sudden peak at late 20 century and before that were used so rarely ($\frac{\sum_{t=1950}^{2000} f_{\text{word}}(t)}{0.99 * \sum_{t=1850}^{2000} f_{\text{word}}(t)} \geq 1$) are deleted.
- The names which have entries in Wikipedia not corresponded to the Russian origin like *Rostow* which refers to Americans or *Romanow* which refers to Polish places are deleted.

For each of the words we count in each year the number of times it occurred with variant 'v', $n_v(i,t)$, and the number of times it occurred with variants 'w' or 'ff', $n_{w+ff}(i,t)$. We consider the timeseries of six representative cases: (i) the average over all words, $\rho_{avg}(t)$; (ii) the five most used words. We calculate the average relative usage of one variant as an average over all tokens, i.e.

$$\rho_{\text{avg}}(t) = \frac{N_v(t)}{N_v(t) + N_{w+ff}(t)} \tag{A.8}$$

with $N_v(t) = \sum_{i=1}^{50} n_v(i,t)$ and $N_{w+ff}(t) = \sum_{i=1}^{50} n_{w+ff}(i,t)$. The respective relative usage for the

words is calculated as follows:

$$\rho_{\text{word}=i}(t) = \frac{n_v(i,t)}{n_v(i,t) + n_{w+ff}(i,t)}. \tag{A.9}$$

The frequency of a word, $f_{\text{word}}(t)$, is measured as the weight of all variants combined over all tokens:

$$f_{\text{word}=i}(t) = \frac{n_{v+w+ff}(i,t)}{\#\text{token}(t)} \tag{A.10}$$

Regularization of verbs

In this section we focus on the regularization of English verbs [Pin99]. In addition to the regular past form of a verb, which is generated by adding -ed (laugh → laughed), there exists a small number of verbs which are conjugated irregularly, e.g. burn → burnt. However, all irregular forms coexist with a corresponding regular variant [MSA+11]. We investigate the competition between the regular and the irregular form for 281 verbs with a recently attested irregular form [MSA+11].

The following filtering is employed. We discard any verb, where the irregular past form is the same as the infinitive since it would not be possible to distinguish between a verb that is used as a past form or a present form, e.g. for the verb 'beat' the irregular preterit is 'beat'. For the remaining verbs we count for each year, $t \in [1800, 2008]$, the number of times it occurs in a regular form, $n_{\text{reg}}(t)$, and the number of times it occurs as an irregular form, $n_{\text{irreg}}(t)$. We condition the counts on those forms that are identified as verbs by the associated part-of-speech tags. As an example, for the verb 'write', $n_{\text{reg}}(t) = n(\text{writed}, t)$, and $n_{\text{irreg}}(t) = n(\text{writ}, t) + n(\text{written}, t) + n(\text{wrote}, t)$, since we have to combine the usage of past participle and preterit to capture all irregular past forms.

From this we can calculate the relative usage of the regular form:

$$\rho(t) = \frac{n_{\text{reg}}(t)}{n_{\text{reg}}(t) + n_{\text{irreg}}(t)}. \tag{A.11}$$

We then select the 10 verbs that exhibit the largest relative change $\mid \rho_1 - \rho_0 \mid$, where ρ_0 and ρ_1 is the average over the 20 datapoints in the beginning ($t \in [1800 - 1819]$) and the end ($t \in [1989 - 2008]$) of the timeseries respectively. These verbs are: abide (abided/abode), burn (burned/burnt), chide (chided/chid,chidden), cleave (cleaved/clove,cloven), light (lighted/lit), smell (smelled/smelt), spell (spelled/spilt), spill (spilled/spilt), thrive (thrived/throve,thriven), wake (waked/woke,waken).

B. Statistical analysis of rank-frequency distribution for different languages

Analysis of different languages

In this section we give a detailed overview of the results obtained from fitting the models in Sec. 3.1.1 to the rank-frequency distributions for all languages considered, i.e., English, French, Spanish, German, and Russian. In Fig. B.1 - B.5(a+b) we plot the AIC from the models in Sec. 3.1.1 applied to yearly $y(t)$ and cumulative data $Y(t)$ of the respective language. In Fig. B.1 - B.5(c) we show explicitly the rank-frequency distribution of the data $Y(2000)$ and the corresponding fits of the three models that yield the best description: the double power law (DP), the power law with an exponential cutoff in the tail (PET), and the log-normal (LN).

For English, DP yields the best description of the yearly data for $t \gtrsim 1950$ and for the cumulative data for $t \gtrsim 1810$. As the databases $y(1950)$ and $Y(1810)$ can be considered independent datasets and by comparing with Fig. A.1(a) we conclude that the size of the database needs to exceed a certain threshold ($\approx 10^9$ tokens) in order to discriminate competing models like PET in the tail. This is further corroborated by looking at the inset in Fig. B.1c), where it can be seen that DP outperforms PET especially in the description of the tail of the distribution.

For the other languages except English the AIC of the yearly data $y(t)$ favors PET. This comes with no surprise since their size is limited to $< 10^9$ tokens for all $t \in [1805, 2000]$, as can be seen in Fig. A.1(a). In contrast, the cumulative data $Y(t)$ shows different results. For French and Spanish the AIC favors DP as the size of the database grows, especially for the largest dataset $Y(2000)$. Again, this becomes clear when looking at the deviations of the fits to the real data in the inset of Fig. B.2(c), B.3(c), which seem to diverge for PET or LN in the tail of the distribution. For German and Russian the AIC identifies DP only as the second best fit for the cumulative data $Y(t)$. This is most probably due to the fact that the size of the database for those languages is still not large enough in order to discriminate a second power law regime clearly. Additionally, for these languages the critical rank b^*, where a transition between the two power laws occurs, is shifted towards higher values, possibly due to the different degree of inflection (see main text). This in turn implies that the fraction of tokens belonging to the power law in the tail is much smaller than in English, which means that a larger database is needed in order to discriminate PET or LN. This claim is further supported by the insets of Fig. B.4(c), B.5(c), where we show that especially in the tail of the distribution DP deviates less from the data than the competing models.

Whereas English, French, and Spanish give approximately the same values for the largest database $Y(2000)$, German and Russian show larger values for b and a different power law exponent in the tail. The latter might point towards more subtle differences between the languages besides inflection.

Wikipedia

In this section we want to show that our findings related to the double power law fit are indeed of general validity and do not originate from peculiarities of the Google-ngram database, e.g. scanning

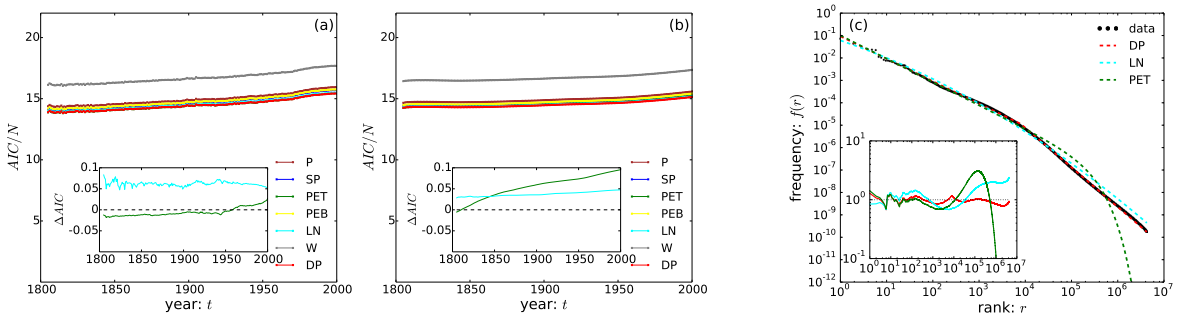

Figure B.1.: Discrimination between different models with AIC for English. Value of the AIC for a) yearly data $y(t)$ b) cumulative data $Y(t)$. The inset shows the difference $\Delta AIC = AIC_{\mathrm{i}}/N - AIC_{\mathrm{DP}}/N$, $i \in \{\mathrm{P}, \mathrm{SP}, \mathrm{PET}, \mathrm{PEB}, \mathrm{LN}, \mathrm{W}\}$ meaning that if $\Delta AIC > 0$ the double power law (DP) is the most likely model among the proposed describing the data. Numbers refer to the indices of the model in Sec. 3.1.1. c) rank-frequency plot for $Y(2000)$ and the fits of the three best models. The inset shows the ratio $f_{\mathrm{data}}(r)/f_{\mathrm{fit}}(r)$.

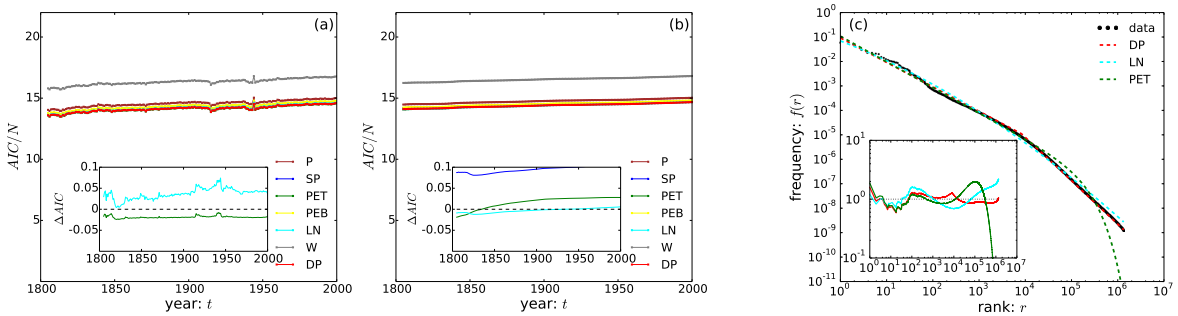

Figure B.2.: Same as in Fig. B.1 for French.

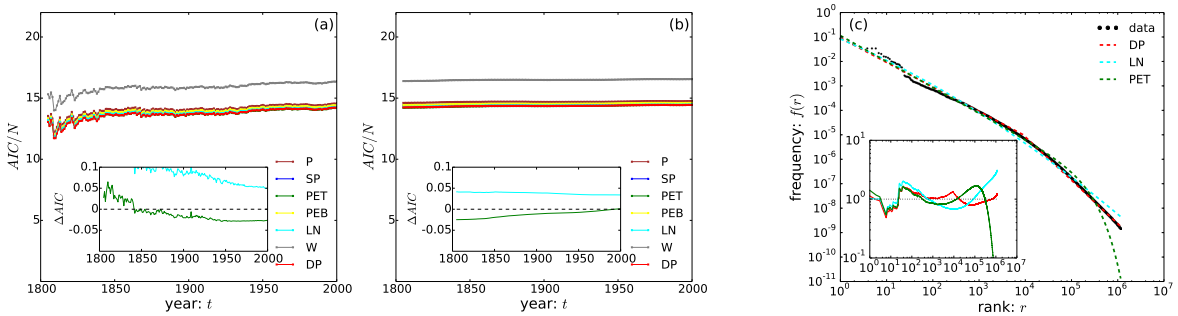

Figure B.3.: Same as in Fig. B.1 for Spanish.

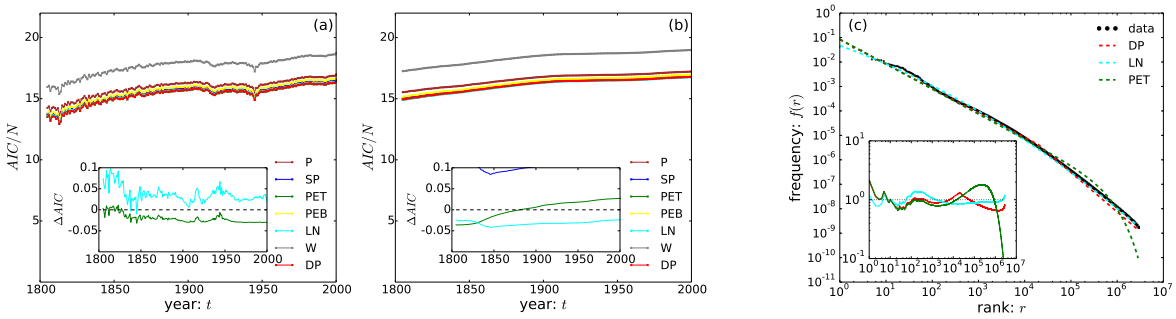

Figure B.4.: Same as in Fig. B.1 for German.

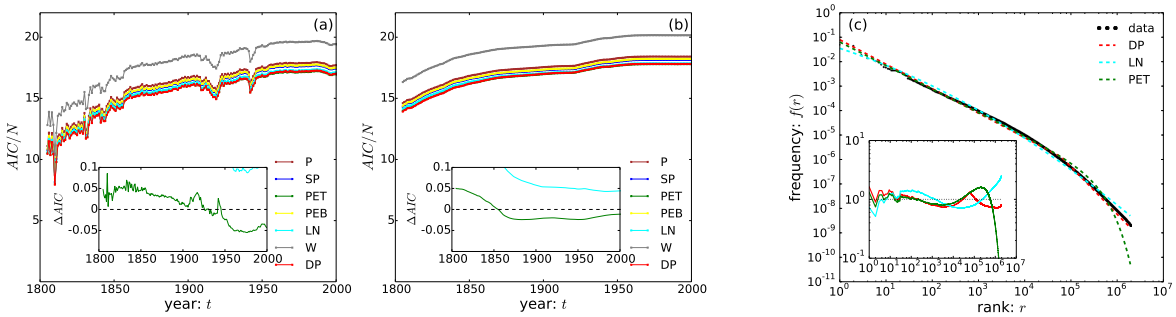

Figure B.5.: Same as in Fig. B.1 for Russian.

problems. For this we choose a complete snapshot of the English Wikipedia [Wik], because i) it contains a large amount of text, ii) the text does not need to be scanned, and iii) the publishing process is inherently different from that of books.

Following the recipe in Sec. 3.1.2, we show that the results for fitting the models inSec. 3.1.1 to the rank-frequency distribution of the Wikipedia database is consistent with the results from the Google-ngram database. In Tab. B.1 we show the values for the AIC from which we can see that the double power law is the best fit among the proposed models with a probability $1-\tilde{l}_{dp} < 10^{-15}$. Additionally, in Fig. B.6 we plot the rank-frequency distribution of the Wikipedia data and the corresponding fits of the three models that yield the best description: the double power law (DP), the power law with an

i	distribution	AIC/N
1	Power law (P)	15.972
2	Shifted power law (SP)	15.782
3	Power law with exponential cutoff, tail (PET)	15.662
4	Power law with exponential cutoff, beginning (PEB)	15.821
5	Log-normal (LN)	15.574
6	Weibull (W)	17.740
7	Double power law (DP)	15.525

Table B.1.: Values AIC/N from fitting the proposed models in Sec. 3.1.1 to the rank-frequency distributions of the English Wikipedia with $N = 1\,257\,349\,755$ tokens.

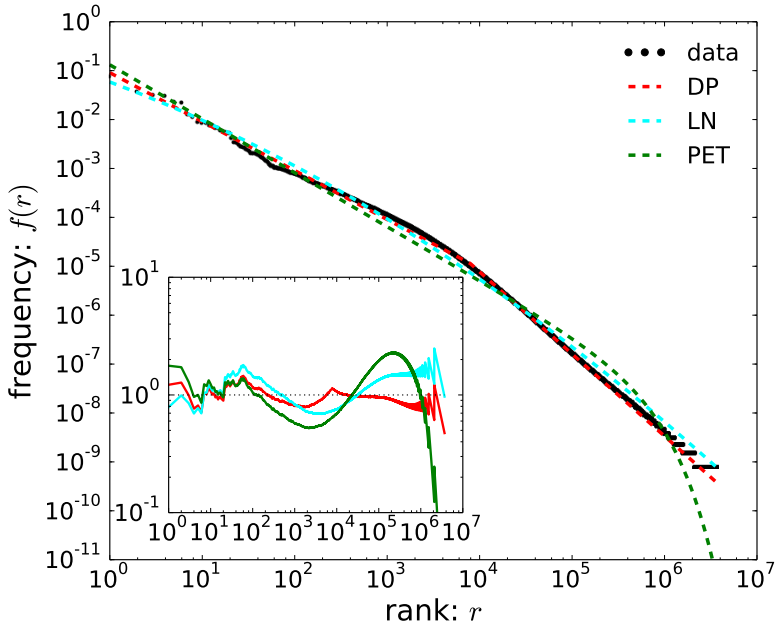

Figure B.6.: Rank frequency distribution for the English Wikipedia and the fits of the three best models. The inset shows the ratio $f_{\text{data}}(r)/f_{\text{fit}}(r)$.

exponential cutoff in the tail (PET), and the log-normal (LN). This corroborates our claim that DP is the best fit for the rank-frequency distribution. Furthermore, the estimated values for the parameters are $\gamma = 1.68$ and $b = 7830$, which closely matches our observations from the Google-ngram database ($\gamma^* = 1.77$, $b^* = 7873$).

C. Rescaling the threshold in Heaps' law

In this section we support findings from Sec. 3.2.1 regarding the vocabulary growth with a threshold s, Eq. (3.31)

$$\mathbb{E}\left[V^{(s)}(N)\right] = \sum_r \left[1 - \sum_{j=0}^{s-1} \frac{(f(r)N)^j}{j!} e^{-f(r)N}\right] \tag{C.1}$$

Looking at the rescaled variable $N \mapsto N/s$ and taking the limit $s \to \infty$ gave the solution, Eq. (3.34),

$$\mathbb{E}\left[V^{(s)}(N' = 1/f(r))\right] = r \tag{C.2}$$

which for the double power law in the rank-frequency distribution, $f(r)$ in Eq. (3.16), with parameters γ and b gives for the vocabulary growth Eq. (3.32)

$$\mathbb{E}\left[V_{\mathrm{dp}}(N;\gamma,b)\right] \approx C_s \begin{cases} N, & N \ll N_b \\ N_b^{1-1/\gamma} N^{1/\gamma}, & N \gg N_b, \end{cases} \tag{C.3}$$

In Fig. C.1 we show the $V^{(s)}(\tilde{N})$ curve obtained from the PNM, Eq. (C.1), for the double power law in the rank-frequency distribution, Eq. (3.16), with parameters $\gamma^* = 1.77$ and $b^* = 7873$ for different thresholds s. One can see that the growth curves for $s > 8$ are almost indistinguishable from the asymptotic solution obtained from Eq. (C.2).

From these observation we conclude that Eq. (C.2) is already a good approximation for $s \gg 1$, where in practice this can mean $s > 10$. As a result we obtain Eq. (C.3) from the main text. This means that the increase of the threshold s leads to a reduction of the fluctuations of the growth curve of the vocabulary and can be explained as a result of a simple stochastic process. In Fig. C.2 we show that this claim holds when applied to real texts of the size of single books, as well as for a collection of several million books, as in Fig. C.3.

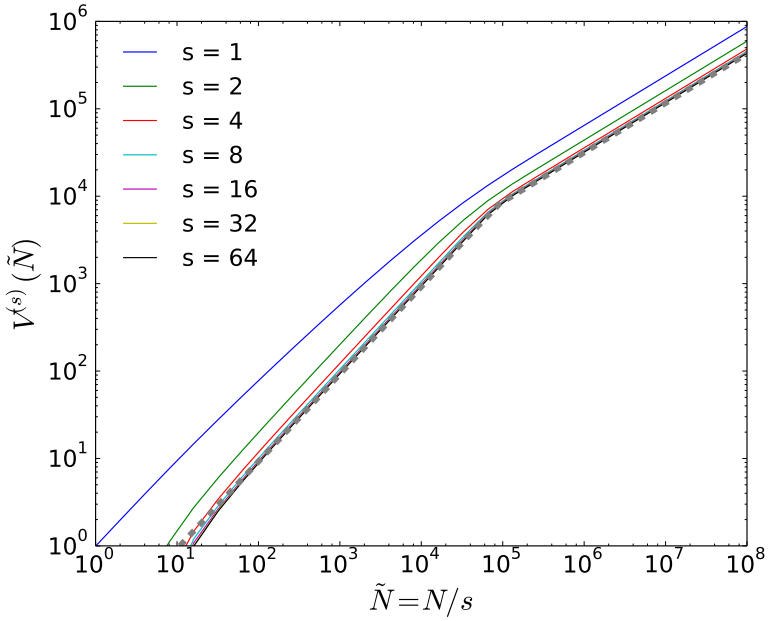

Figure C.1.: Influence of threshold s on size of vocabulary for the PNM. Growth curves $V^{(s)}(\tilde{N} = N/s)$ obtained from PNM, Eq. (C.1), for double power law, Eq. (3.16), with parameters $\gamma^* = 1.77$, $b^* = 7873$ with different thresholds s. The dashed curve shows the asymptotic solution Eq. (C.3).

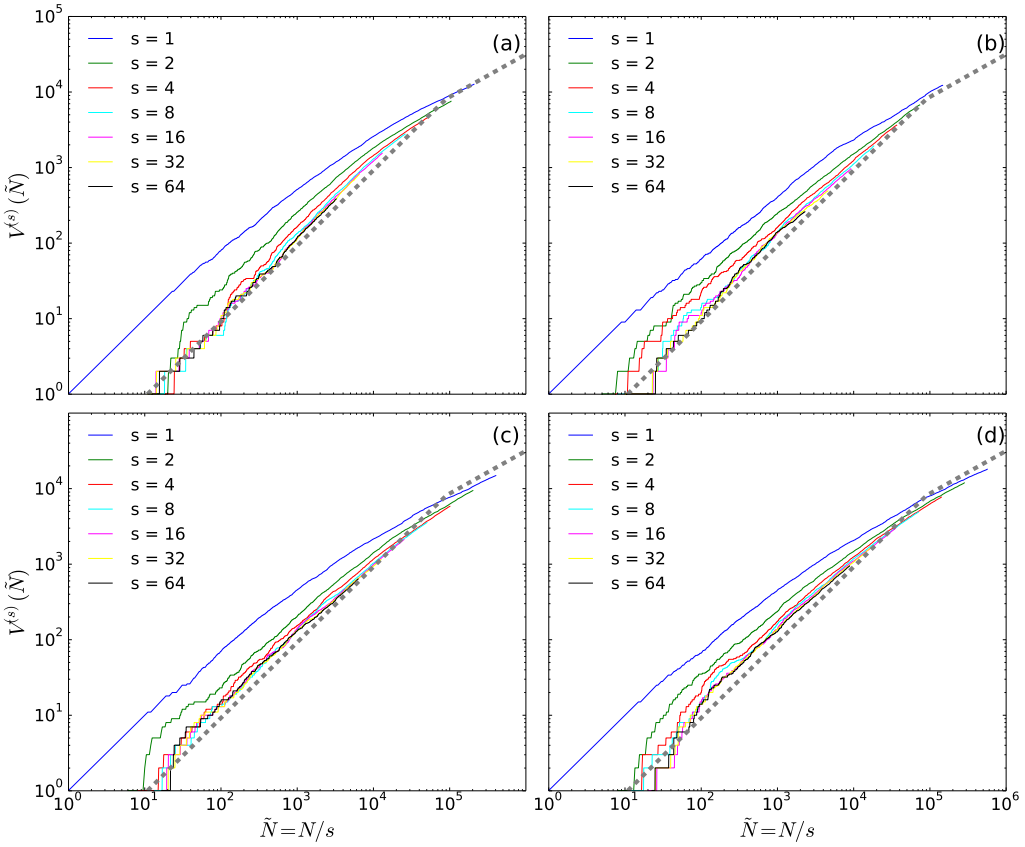

Figure C.2.: Influence of threshold s on the size of the vocabulary for single books. Growth curves $V^{(s)}(\tilde{N} = N/s)$ obtained from 4 different books with different thresholds s. a) Charles Darwin: "The Voyage of the Beagle" b) Mark Twain: "Life on the Mississippi" c) Miguel de Cervantes Saavedra: "Don Quixote", translated by John Ormsby d) Leo Tolstoy: "War and Peace", translated by Louise and Aylmer Maude. All texts were retrieved from the Project Gutenberg (www.gutenberg.org) on 21.09.2010. The dashed curve shows the asymptotic solution Eq. (C.3) of the PNM assuming a double power law, Eq. (3.16), with parameters $\gamma^* = 1.77$ and $b^* = 7873$.

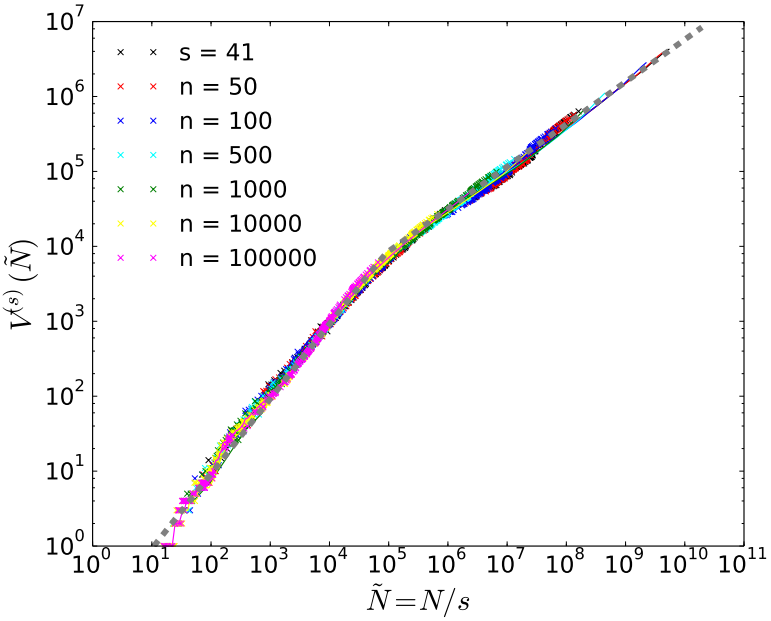

Figure C.3.: Influence of threshold s on size of vocabulary for the English Google-ngram database. Growth curves $V^{(s)}(\tilde{N} = N/s)$ obtained from yearly data $y(t)$ (x-symbol) and cumulative data $Y(t)$ (line) for different values of the threshold s with $s \in [41, 10^5]$ marked by different colors. The dashed curve shows the asymptotic solution Eq. (C.3) of the PNM assuming a double power law, Eq. (3.16), with parameters $\gamma^* = 1.77$, $b^* = 7873$.

D. JSD-α with weights

Here we discuss how to proceed if D_α is computed from finite datasets with different finite lengths N, i.e. when $\boldsymbol{p}$ ($\boldsymbol{q}$) is estimated from a sequence of length N_p ($N_q \neq N_p$).

D.1. Different weights

A possible way to extend Eq. (4.30) taking into account the unequal contribution $N_p \neq N_q$ is to consider weights π as [GBGC+02]

$$D_\alpha^\pi(\boldsymbol{p},\boldsymbol{q}) = H_\alpha(\pi_p\boldsymbol{p} + \pi_q\boldsymbol{q}) - \pi_p H_\alpha(\boldsymbol{p}) - \pi_q H_\alpha(\boldsymbol{q}). \tag{D.1}$$

with $\pi_p = N_p/N$ and N_q/N such that $\pi_p + \pi_q = 1$ with $N = N_p + N_q$ (denoted as natural weights in the following). Obviously, if $N_p = N_q$ then $\pi_p = \pi_q = 1/2$ and D_α is recovered. The normalized distance (4.33) becomes

$$\tilde{D}_\alpha^\pi(\boldsymbol{p},\boldsymbol{q}) = \frac{D_\alpha^\pi(\boldsymbol{p},\boldsymbol{q})}{D_\alpha^{\pi,\max}(\boldsymbol{p},\boldsymbol{q})}, \tag{D.2}$$

where

$$D_\alpha^{\pi,\max}(\boldsymbol{p},\boldsymbol{q}) = \left(\pi_p^\alpha - \pi_p\right) H_\alpha\left(\boldsymbol{p}\right) + \left(\pi_q^\alpha - \pi_q\right) H_\alpha\left(\boldsymbol{q}\right) + \frac{1}{1-\alpha}\left(\pi_p^\alpha + \pi_q^\alpha - 1\right). \tag{D.3}$$

Our main results for the finite-size scaling of D_α summarized in Tab. 4.2 remain valid for the weighted divergences.

The approach above follows Ref [GBGC+02], which introduced weights to the usual JSD (non-normalized, $\alpha = 1$) and showed that the natural weights $\pi_p = N_p/N$ and $\pi_q = N_q/N$ imply certain useful properties for the JSD, e.g., that the bias does not depend on the relative size of the two samples. While their main motivation was to compare the statistical significance of a single measurement of the JSD in the identification of stationary subsequences (of possibly different lengths) in a non-stationary symbolic sequence, here, we are mainly interested in comparing two (or more) measured distances. In this case, choosing weights that depend on the size of the individual samples becomes problematic when the sequences are of different lengths. The demonstration that $\sqrt{D_\alpha(\boldsymbol{p},\boldsymbol{q})}$ is a metric for any $\alpha \in (0,2]$ [BH09] is valid for fixed weights $\pi_p = \pi_q = 1/2$. More generally, the measure D_α^π itself depends on the weights π such that D_α^π and $D_\alpha^{\pi'}$ constitute different measures when $\pi \neq \pi'$. It is therefore not meaningful to compare $D_\alpha^\pi(\boldsymbol{p},\boldsymbol{q})$ and $D_\alpha^{\pi'}(\boldsymbol{p}',\boldsymbol{q}')$ if $N_p/N_q \neq N_{p'}/N_{q'}$ because this would imply that $\pi' \neq \pi$.

D.2. Equal weights

In the previous section we argued that it is essential to choose fixed weights π when comparing different distances. The choice of equal weights $\pi_p = \pi_q = 1/2$ can, however, still be interpreted in the framework of natural weights ($\pi_p = N_p/N$, $\pi_q = N_q/N$) as the distance between under-sampled

versions of the sequences. For given $\boldsymbol{p}$ and $\boldsymbol{q}$ with $N_p \neq N_q$ we choose equal weights $\pi_p = \pi_q = 1/2$ yielding a distance $D_\alpha^{1/2}(\boldsymbol{p}, \boldsymbol{q})$. If we randomly draw samples $\boldsymbol{p}'$ and $\boldsymbol{q}'$ from the distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ the equal weights of size $N'_p = N'_q$, by construction the natural weights coincide with the equal weights, i.e. $\pi'_p = \pi'_q = N'_p/N = N'_q/N = 1/2$, and $\lim_{N'_p = N'_q \to \infty} D_\alpha^{\pi'}(\boldsymbol{p}', \boldsymbol{q}') = D_\alpha^{1/2}(\boldsymbol{p}, \boldsymbol{q})$.

In Fig. D.1 we show the difference in $\tilde{D}_\alpha^{\pi}(\boldsymbol{p}, \boldsymbol{q})$ between two empirical distributions from the Google-ngram with different sizes ($N_p \neq N_q$) when choosing equal and natural weights. Using equal weights corresponds to the case in which we draw samples $\hat{\boldsymbol{p}}$ and $\hat{\boldsymbol{q}}$ that are of equal length ($N'_p = N'_q$) such that equal and natural weights coincide and taking the limit $N'_p, N'_q \to \infty$.

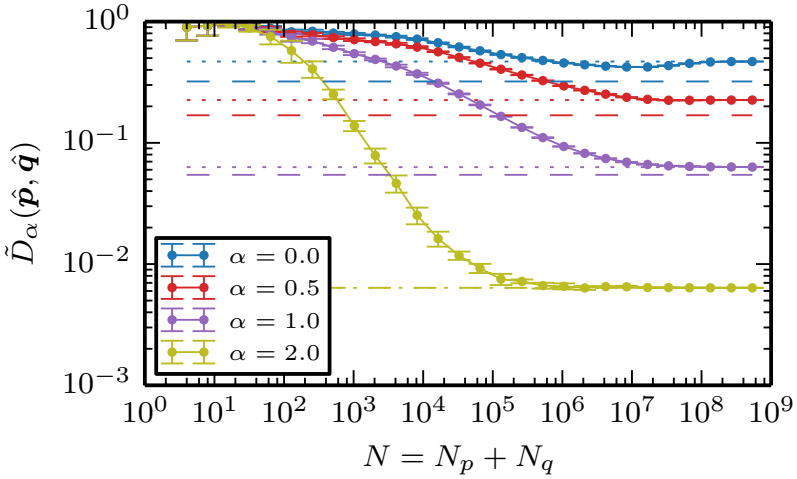

Figure D.1.: JSD-α for sequences of different lengths. Measurement of $\tilde{D}_\alpha(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})$ between sequences $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{q}}$ of size $N'_p = N'_q$ sampled randomly from the empirical distribution of the Google-ngram of the years $t \in \{1850, 1950\}$ with different sizes, i.e. $\boldsymbol{p} = \boldsymbol{p}_{t=1850}$ and $\boldsymbol{q} = \boldsymbol{p}_{t=1950}$ with $N_p \neq N_q$, as a function of the sample size $N' = N'_p + N'_q$ for different values of α. The dotted (dashed) lines show $\tilde{D}_\alpha^{\pi}(\boldsymbol{p}, \boldsymbol{q})$ between the full distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ with equal (natural) weights, i.e. $\pi_p = \pi_q = 1/2$ ($\pi_p = N_p/(N_p + N_q) \approx 0.22$ and $\pi_q = N_q/(N_p + N_q) \approx 0.78$ corresponding to the relative size of $\boldsymbol{p}$ and $\boldsymbol{q}$)

Bibliography

[AB02] R. Albert and A.-L. Barabási, *Statistical mechanics of complex networks*, Rev. Mod. Phys. **74** (2002), 47–97. Cited on pages 2 and 67.

[AB04] M. Argollo de Menezes and A.-L. Barabási, *Separating Internal and External Dynamics of Complex Systems*, Phys. Rev. Lett. **93** (2004), 068701. Cited on page 89.

[ACE12] E.G. Altmann, G. Cristadoro, and M.D. Esposti, *On the origin of long-range correlations in texts*, Proc. Nat. Acad. Sci. USA **109** (2012), 11582–7. Cited on pages 3, 11, 23, 113, and 115.

[AG16] E.G. Altmann and M. Gerlach, *Statistical laws in linguistics*, Creativity and Universality in Language (M.D. Esposti, E.G. Altmann, and F. Pachet, eds.), Springer International Publishing, 2016. Cited on pages 5 and 15.

[Ait01] J. Aitchison, *Language Change: Progress Or Decay?*, Cambridge University Press, Cambridge, 2001. Cited on page 83.

[AJK06] S. Albeverio, V. Jentsch, and H. Kantz, *Extreme Events in Nature and Society*, Springer Verlag, Berlin Heidelberg, 2006. Cited on pages 8 and 9.

[Aka74] H. Akaike, *A new look at the statistical model identification*, IEEE T. Automat. Contr. **19** (1974), 716–723. Cited on page 18.

[ALDEM06] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, and E. Moses, *Hierarchical structures induce long-range dynamical correlations in written texts*, Proc. Nat. Acad. Sci. USA **103** (2006), 7956–61. Cited on page 3.

[Alt80] G. Altmann, *Prolegomena to Menzerath's law*, Glottometrika **2** (1980), 1. Cited on page 11.

[And72] P.W. Anderson, *More is different*, Science **177** (1972), 393–6. Cited on pages 1 and 7.

[API] Plos API, *All articles published in the journal plos one*, `http://api.plos.org/` [Online; accessed 17-October-2013]. Cited on pages 13 and 119.

[APM09] E.G. Altmann, J.B. Pierrehumbert, and A.E. Motter, *Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words*, PloS one **4** (2009), e7678. Cited on pages 11, 23, 35, and 113.

[APM11] ———, *Niche as a determinant of word fate in online groups*, PloS one **6** (2011), e19009. Cited on pages 35 and 37.

[APP14] E. Archer, I.M. Park, and J.W. Pillow, *Bayesian entropy estimation for countable discrete distributions*, J. Mach. Learn. Res. **15** (2014), 2833–2868. Cited on page 59.

[Arr21] O. Arrhenius, *Species and Area*, J. Ecol. **9** (1921), 95. Cited on pages 7 and 24.

[AS72] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions*, Dover Publications, New York, 1972. Cited on pages 26 and 45.

[Aue13] F. Auerbach, *Das Gesetz der Bevölkerungskonzentration*, Petermanns Geographische Mitteilungen **59** (1913), 73–76. Cited on pages 1 and 7.

[BA99] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science **286** (1999), 509–12. Cited on pages 2 and 68.

[BA02] K.P. Burnham and D.R. Anderson, *Model Selection and Multimodel Inference*, Springer-Verlag, New York, 2002. Cited on pages 18, 94, and 95.

[Baa01] R.H. Baayen, *Word Frequency Distributions*, Kluwer Academic Publishers, Dordrecht, Netherlands, 2001. Cited on pages 11, 16, 23, 24, and 45.

[Bal04] P. Ball, *Critical Mass: How One Thing Leads to Another*, Farrar, Straus and Giroux, London, 2004. Cited on page 1.

[Bas59] G.P. Basharin, *On a Statistical Estimate for the Entropy of a Sequence of Independent Random Variables*, Theor. Probab. Appl. **4** (1959), 333–336. Cited on page 52.

[Bas69] F.M. Bass, *A New Product Growth for Model Consumer Durables*, Manage. Sci. **15** (1969), 215–227. Cited on pages 88, 91, and 106.

[Bas04] ———, *Comments on "A New Product Growth for Model Consumer Durables The Bass Model"*, Manage. Sci. **50** (2004), 1833–1840. Cited on pages 88 and 106.

[Bat13] M. Batty, *Sociology. A theory of city size*, Science **340** (2013), 1418–9. Cited on page 8.

[Bau07] H. Bauke, *Parameter estimation for power-law distributions by maximum likelihood methods*, Eur. Phys. J. B **58** (2007), 167–173. Cited on page 17.

[BBB+09] C. Beckner, R. Blythe, J. Bybee, M.H. Christiansen, W. Croft, N.C. Ellis, J. Holland, J. Ke, D. Larsen-Freeman, and T. Schoenemann, *Language Is a Complex Adaptive System: Position Paper*, Lang. Learn. **59** (2009), 1–26. Cited on pages 2 and 3.

[BBCM06] G. Baxter, R. Blythe, W. Croft, and A. McKane, *Utterance selection model of language change*, Phys. Rev. E **73** (2006). Cited on pages 3, 88, and 114.

[BBCM09] ———, *Modeling language change: An evaluation of Trudgill's theory of the emergence of New Zealand English*, Lang. Var. Change **21** (2009), 257. Cited on page 3.

[BBM11] S.K. Baek, S. Bernhardsson, and P. Minnhagen, *Zipf's law unzipped*, New J. Phys. **13** (2011), 043004. Cited on page 16.

[BC12] R. Blythe and W. Croft, *S-curves and the mechanisms of propagation in language change*, Language **88** (2012), 269–304. Cited on pages 3, 87, 88, and 115.

[BCAKCC06] L.M.A. Bettencourt, A. Cintrón-Arias, D.I. Kaiser, and C. Castillo-Chávez, *The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models*, Physica A **364** (2006), 513–536. Cited on page 88.

[BCM09] S. Bernhardsson, L.E. Correa da Rocha, and P. Minnhagen, *The meta book and size-dependent properties of written language*, New J. Phys. **11** (2009), 123015. Cited on pages 24 and 25.

[BCN08] M. Beltrán del Río, G. Cocho, and G.G. Naumis, *Universality in the tail of musical note rank distribution*, Physica A **387** (2008), 5552–5560. Cited on page 48.

[BDO95] M.W. Berry, S.T. Dumais, and G.W. O'Brien, *Using Linear Algebra for Intelligent Information Retrieval*, SIAM Rev. **37** (1995), 573–595. Cited on page 64.

[BFPS+13] A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M.H. Christiansen, *Networks in cognitive science*, Trends Cogn. Sci. **17** (2013), 348–60. Cited on page 11.

[BG90] J.-P. Bouchaud and A. Georges, *Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications*, Rev. Mod. Phys. **195** (1990), 127–293. Cited on pages 8 and 58.

[BGGC+00] P. Bernaola-Galván, I. Grosse, P. Carpena, J.L. Oliver, R. Román-Roldán, and H.E. Stanley, *Finding borders between coding and noncoding DNA regions by an entropic segmentation method*, Phys. Rev. Lett. **85** (2000), 1342–1345. Cited on page 59.

[BGOB12] R.A. Bentley, P. Garnett, M. J. O'Brien, and W.A. Brock, *Word diffusion and climate science*, PloS one **7** (2012), e47966. Cited on page 106.

[BH09] J. Briët and P. Harremoës, *Properties of classical and quantum Jensen-Shannon divergence*, Phys. Rev. A **79** (2009), 052311. Cited on pages 50 and 131.

[Bib] Bibliography, *Physicists' papers on natural language from a complex systems viewpoint*, `http://www.pks.mpg.de/~edugalt/physicist-language/` [Online; accessed 20-September-2015]. Cited on page 3.

[BKN11] B. Ball, B. Karrer, and M.E.J. Newman, *Efficient and principled method for detecting communities in networks*, Phys. Rev. E **84** (2011), 036103. Cited on page 68.

[Ble12] D.M. Blei, *Probabilistic topic models*, Commun. ACM **55** (2012), 77. Cited on pages 38 and 66.

[BLH+07] L.M.A. Bettencourt, J. Lobo, D. Helbing, C. Kühnert, and G.B. West, *Growth, innovation, scaling, and the pace of life in cities*, Proc. Nat. Acad. Sci. USA **104** (2007), 7301–6. Cited on page 8.

[BLNZ95] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A Limited Memory Algorithm for Bound Constrained Optimization*, SIAM J. Sci. Comput. **16** (1995), 1190–1208. Cited on page 103.

[BLT12] A. Baronchelli, V. Loreto, and F. Tria, *Language Dynamics*, Adv. Complex Syst. **15** (2012), 1203002–1. Cited on page 88.

[BnCPS13] M. Boguñá, C. Castellano, and R. Pastor-Satorras, *Nature of the Epidemic Threshold for the Susceptible-Infected-Susceptible Dynamics in Networks*, Phys. Rev. Lett. **111** (2013), 068701. Cited on page 114.

[BND+11] K.W. Boyack, D. Newman, R.J. Duhon, R. Klavans, M. Patek, J.R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, and K. Börner, *Clustering more than two million biomedical publications: comparing the accuracies of nine text-based similarity approaches*, PloS one **6** (2011), e18029. Cited on page 49.

[BNJ03] D.M. Blei, A.Y. Ng, and M.I. Jordan, *Latent Dirichlet Allocation*, J. Mach. Learn. Res. **3** (2003), 993–1022. Cited on pages 38, 43, 65, 66, and 67.

[Box79] G.E.P. Box, *Robustness in the strategy of scientific model building*, Robustness in statistics, Academic Press, New York, 1979. Cited on page 2.

[BR82] J. Burbea and C. Rao, *On the convexity of some divergence measures based on entropy functions*, IEEE T. Inform. Theory **28** (1982), 489–495. Cited on page 49.

[BR85] R. Boyd and P.J. Richerson, *Culture and the Evolutionary Process*, University of Chicago Press, Chicago, IL, 1985. Cited on pages 88 and 89.

[Bra82] B. Brainerd, *On the Relation between the Type-Token and Species-Area Problems*, J. Appl. Probab. **19** (1982), 785–793. Cited on pages 24 and 115.

[BSW14] V. Bochkarev, V. Solovyev, and S. Wichmann, *Universals versus historical contingencies in lexical evolution*, J. R. Soc. Interface **11** (2014), 20140841. Cited on page 49.

[BTW87] P. Bak, C. Tang, and K. Wiesenfeld, *Self-organized criticality: An explanation of the 1/f noise*, Phys. Rev. Lett. **59** (1987), 381–384. Cited on pages 1 and 8.

[BYN00] R. Baeza-Yates and G. Navarro, *Block addressing indices for approximate text retrieval*, J. Am. Soc. Inform. Sci. **51** (2000), 69–82. Cited on pages 24 and 25.

[Can] Canoonet, *German dictionaries and grammar*, `http://www.canoo.net` [Online; accessed 03-April-2013]. Cited on page 120.

[CB11] R. Chicheportiche and J.-P. Bouchaud, *Goodness-of-Fit tests with Dependent Observations*, J. Stat. Mech. **2011** (2011), P09003. Cited on page 23.

[CB15] D. Centola and A. Baronchelli, *The spontaneous emergence of conventions: An experimental study of cultural evolution*, Proc. Nat. Acad. Sci. USA **112** (2015), 1989–1994. Cited on page 3.

[CBB+09] C. Cattuto, A. Barrat, A. Baldassarri, G. Schehr, and V. Loreto, *Collective dynamics of social annotation*, Proc. Nat. Acad. Sci. USA **106** (2009), 10511–5. Cited on pages 24 and 115.

[CFL09] C. Castellano, S. Fortunato, and V. Loreto, *Statistical physics of social dynamics*, Rev. Mod. Phys. **81** (2009), 591–646. Cited on pages 2 and 88.

[CG95] K.W. Church and W.A. Gale, *Poisson Mixtures*, Nat. Lang. Eng. **1** (1995), 163–190. Cited on pages 35, 36, and 44.

[CL14] J. Cong and H. Liu, *Approaching human language with complex networks*, Phys. Life Rev. **11** (2014), 598–618. Cited on page 11.

[CM09] M. Choudhury and A. Mukherjee, *The Structure and Dynamics of Linguistic Networks*, Dynamics On and Of Complex Networks (N. Ganguly, A. Deutsch, and A. Mukherjee, eds.), Birkhäuser, Boston, MA, 2009. Cited on page 11.

[CMFS11] B. Corominas-Murtra, J. Fortuny, and R.V. Solé, *Emergence of Zipf's law in the evolution of communication*, Phys. Rev. E **83** (2011), 036115. Cited on pages 3 and 29.

[CMH97] A. Cohen, R.N. Mantegna, and S. Havlin, *Numerical Analysis of Word Frequencies in Artificial and Natural Language Texts*, Fractals **05** (1997), 95–104. Cited on page 16.

[CMS09] B. Croft, D. Metzler, and T. Strohmann, *Search Engines: Information Retrieval in Practice*, Addison-Wesley, Boston, MA, 2009. Cited on pages 2 and 24.

[CPC+14] C.F. Cuskley, M. Pugliese, C. Castellano, F. Colaiori, V. Loreto, and F. Tria, *Internal and external dynamics in language: evidence from verb regularity in a historical corpus of english*, PloS one **9** (2014), e102882. Cited on page 103.

[Cra05] I. Cramer, *The Parameters of the Altmann-Menzerath Law*, J. Quant. Linguist. **12** (2005), 41–52. Cited on page 11.

[Cro00] W. Croft, *Explaining language change: an evolutionary approach*, Pearson Education, Essex, 2000. Cited on pages 3 and 83.

[CS08] R. Crane and D. Sornette, *Robust dynamic classes revealed by measuring the response function of a social system*, Proc. Nat. Acad. Sci. USA **105** (2008), 15649–53. Cited on page 89.

[CSN09] A. Clauset, C.R. Shalizi, and M.E.J. Newman, *Power-Law Distributions in Empirical Data*, SIAM Rev. **51** (2009), 661. Cited on pages 7, 8, 17, 18, 22, 23, 113, and 114.

[CT06] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 2006. Cited on pages 12 and 49.

[D'A86] R.B. D'Agostino, *Goodness-of-Fit-Techniques*, Marcel Dekker, New York, 1986. Cited on page 19.

[DC13] A. Deluca and Á. Corral, *Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions*, Acta Geophys. **61** (2013), 1351–1394. Cited on page 8.

[DDF+90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, *Indexing by latent semantic analysis*, J. Am. Soc. Inform. Sci. **41** (1990), 391–407. Cited on page 64.

[Deb06] Ł. Debowski, *On Hilberg's law and its links with Guiraud's law**, J. Quant. Linguist. **13** (2006), 81–109. Cited on page 11.

[DKMZ11a] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E **84** (2011), 1–19. Cited on page 114.

[DKMZ11b] ———, *Inference and Phase Transitions in the Detection of Modules in Sparse Networks*, Phys. Rev. Lett. **107** (2011), 065701. Cited on pages 72, 78, and 116.

[DLM] *NIST Digital Library of Mathematical Functions*, `http://dlmf.nist.gov/`, Release 1.0.9 of 2014-08-29, Online companion to [OLBC10]. Cited on pages 27 and 145.

[DM73] F.J. Damerau and B. Mandelbrot, *Tests of the degree of word clustering in samples of written English*, Linguistics **102** (1973), 58–75. Cited on pages 3, 11, and 23.

[DMA12] R. Dickman, N.R. Moloney, and E.G. Altmann, *Analysis of an information-theoretic model for communication*, J. Stat. Mech. **2012** (2012), P12022. Cited on page 3.

[dW99] T.D. de Wit, *When do finite sample effects significantly affect entropy estimates?*, Eur. Phys. J. B **11** (1999), 513–516. Cited on page 52.

[EBK08] Z. Eisler, I. Bartos, and J. Kertész, *Fluctuation scaling in complex systems: Taylor's law and beyond*, Adv. Phys. **57** (2008), 89–142. Cited on pages 39 and 40.

[Eec04] J. Eeckhout, *Gibrat's Law for (All) Cities*, Am. Econ. Rev. **94** (2004), 1429–1451. Cited on page 8.

[Eec09] ———, *Gibrat's Law for (All) Cities: Reply*, Am. Econ. Rev. **99** (2009), 1676–1683. Cited on page 8.

[Egg07] L. Egghe, *Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments*, J. Am. Soc. Inform. Sci. Tec. **58** (2007), 702–709. Cited on page 11.

[Eli11] I. Eliazar, *The growth statistics of Zipfian ensembles: Beyond Heaps' law*, Physica A **390** (2011), 3189–3203. Cited on pages 24 and 25.

[EP94] W. Ebeling and T. Pöschel, *Entropy and Long-Range Correlations in Literary English*, Europhys. Lett. **26** (1994), 241–246. Cited on pages 3 and 11.

[ES03] D.M. Endres and J.E. Schindelin, *A new metric for probability distributions*, IEEE T. Inform. Theory **49** (2003), 1858–1860. Cited on page 49.

[Est16] J.B. Estoup, *Les Gammes Sténographiques*, Institut Sténographique de France, Paris, 1916. Cited on pages 1 and 7.

[FCBC13] F. Font-Clos, G. Boleda, and Á. Corral, *A scaling law beyond Zipf's law and its relation to Heaps' law*, New J. Phys. **15** (2013), 093033. Cited on page 24.

[FCC15] F. Font-Clos and Á. Corral, *Log-Log Convexity of Type-Token Growth in Zipf's Systems*, Phys. Rev. Lett. **114** (2015), 1–13. Cited on page 24.

[Fel68] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume I*, Wiley, New York, 1968. Cited on page 30.

[Fer05] R. Ferrer-i-Cancho, *Zipf's law from a communicative phase transition*, Eur. Phys. J. B **47** (2005), 449–457. Cited on page 16.

[FK03] C. Furusawa and K. Kaneko, *Zipf's Law in Gene Expression*, Phys. Rev. Lett. **90** (2003), 088102. Cited on pages 48 and 116.

[For10] S. Fortunato, *Community detection in graphs*, Phys. Rep. **486** (2010), 75–174. Cited on page 67.

[FS01] R. Ferrer-i-Cancho and R.V. Solé, *Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf?s Law Revisited*, J. Quant. Linguist. **8** (2001), 165–173. Cited on pages 20, 29, and 111.

[FS03] R. Ferrer-i-Cancho and R.V. Sole, *Least effort and the origins of scaling in human language*, Proc. Nat. Acad. Sci. USA **100** (2003), 788–91. Cited on page 3.

[GA13] M. Gerlach and E.G Altmann, *Stochastic Model for the Vocabulary Growth in Natural Languages*, Phys. Rev. X **3** (2013), 021006. Cited on pages 5, 15, 24, 60, 83, and 111.

[GA14] M. Gerlach and E.G. Altmann, *Scaling laws and fluctuations in the statistics of word frequencies*, New J. Phys. **16** (2014), 113010. Cited on pages 5, 11, 35, 60, and 111.

[GBGC+02] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. Stanley, *Analysis of symbolic sequences using the Jensen-Shannon divergence*, Phys. Rev. E **65** (2002), 041905. Cited on pages 38, 49, 59, and 131.

[GCSR03] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis*, CRC Press, Boca Raton, FL, 2003. Cited on page 66.

[GFCA15] M. Gerlach, F. Font-Clos, and E.G. Altmann, *On the similarity of symbol frequency distributions with heavy tails*, arXiv:1510.00277 (2015). Cited on pages 5, 35, 83, and 112.

[GG06] H. García Martín and N. Goldenfeld, *On the origin and robustness of power-law species-area relationships in ecology*, Proc. Nat. Acad. Sci. USA **103** (2006), 10310–5. Cited on page 24.

[GGMA14] F. Ghanbarnejad, M. Gerlach, J.M. Miotto, and E.G. Altmann, *Extracting information from S-curves of language change*, J. R. Soc. Interface **11** (2014), 20141044. Cited on pages 5, 83, and 113.

[Gle11] J.P. Gleeson, *High-Accuracy Approximation of Binary-State Dynamics on Networks*, Phys. Rev. Lett. **107** (2011), 068701. Cited on pages 91 and 92.

[Gle13] ———, *Binary-State Dynamics on Complex Networks: Pair Approximation and Beyond*, Phys. Rev. X **3** (2013), 021004. Cited on pages 91 and 98.

[GLSW96] R. Günther, L. Levitin, B. Schapiro, and P. Wagner, *Zipf 's law and the effect of ranking on probability distributions*, Int. J. Theor. Phys. **35** (1996), 395–417. Cited on page 22.

[GMY04] M.L. Goldstein, S.A. Morris, and G.G. Yen, *Problems with fitting to the power-law distribution*, Eur. Phys. J. B **41** (2004), 255–258. Cited on page 17.

[GPA15] M. Gerlach, T.P. Peixoto, and E.G. Altmann, *Topic models and stochastic block models*, in preparation. Cited on pages 5, 63, and 112.

[Gr7] P.D. Grünwald, *The Minimum Description Length Principle*, MIT Press, Cambridge, MA, 2007. Cited on pages 12, 69, and 74.

[Gra89] P. Grassberger, *Estimating the information content of symbol sequences and efficient codes*, IEEE T. Inform. Theory **35** (1989), 669–675. Cited on page 3.

[Gra08] ———, *Entropy Estimates from Insufficient Samplings*, arXiv:physics/0307138 (2008). Cited on page 59.

[GS04] T.L. Griffiths and M. Steyvers, *Finding scientific topics*, Proc. Nat. Acad. Sci. USA **101 Suppl** (2004), 5228–35. Cited on page 66.

[GSP09] R. Guimerà and M. Sales-Pardo, *Missing and spurious interactions and the reconstruction of complex networks*, Proc. Nat. Acad. Sci. USA **106** (2009), 22073–8. Cited on page 68.

[HAF+10] N. Hend, M. Aerts, C. Faes, Z. Shkedy, O. Lejeune, P. van Damme, and P. Beutels, *Seventy-five years of estimating the force of infection from current status data*, Epidemiol. and Infect. **138** (2010), 802. Cited on page 91.

[Har75] B. Harris, *The Statistical Estimation of Entropy in the non-parametric case*, Colloq. Math. Soc. J. B **16** (1975), 323–355. Cited on page 52.

[HC67] J. Havrda and F. Charvát, *Quantification method of classification processes: Concept of structural a-entropy*, Kybernetika **3** (1967), 30–35. Cited on page 49.

[HCB+09] D.J. Hruschka, M.H. Christiansen, R.A. Blythe, W. Croft, P. Heggarty, S.S. Mufwene, J.B. Pierrehumbert, and S. Poplack, *Building social cognitive models of language change*, Trends Cogn. Sci. **13** (2009), 464–9. Cited on page 103.

[Hea78] H.S. Heaps, *Information Retrieval*, Academic Press, New York, 1978. Cited on pages 11, 24, and 52.

[Her58] G. Herdan, *The relation between the dictionary distribution and the occurrence distribution of word length and its importance fo the study of quantitative linguistics*, Biometrika **45** (1958), 222–228. Cited on page 24.

[Her60] ———, *Type-token mathematics*, Mouton, Den Haag, 1960. Cited on page 52.

[Her64] ———, *Quantitative Linguistics*, Butterworth Press, Oxford, 1964. Cited on page 11.

[HLL83] P.W. Holland, K.B. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Soc. Networks **5** (1983), 109–137. Cited on page 68.

[Hof99] T. Hofmann, *Probabilistic latent semantic indexing*, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '99 (New York), ACM, 1999. Cited on page 65.

[Hof01] ———, *Unsupervised Learning by Probabilistic Latent Semantic Analysis*, Mach. Learn. **42** (2001), 177–196. Cited on page 65.

[HRN12] D. Hu, P. Ronhovde, and Z. Nussinov, *Phase transitions in random Potts systems and the community detection problem: spin-glass type and dynamic perspectives*, Philos. Mag. **92** (2012), 406–445. Cited on page 72.

[HSE94] H. Herzel, A.O. Schmitt, and W. Ebeling, *Finite sample effects in sequence analysis*, Chaos Soliton Fract. **4** (1994), 97–113. Cited on pages 52, 55, and 59.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2009. Cited on pages 17, 85, 94, 95, 98, and 103.

[J+10] F. Johansson et al., *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 0.14)*, February 2010, `http://code.google.com/p/mpmath/`. Cited on page 17.

[JÏ2] G. Jäger, *Power laws and other heavy-tailed distributions in linguistic typology*, Adv. Complex Syst. **15** (2012), 1150019–1. Cited on page 16.

[Joh05] S. Johnson, *Spelling trouble: Language, ideology and the reform of German orthography*, Multilingual Matters, Clevedon, UK, 2005. Cited on pages 102 and 120.

[JOP+] E. Jones, T. Oliphant, P. Peterson, et al., *SciPy: Open source scientific tools for Python*, 2001–, `http://www.scipy.org`. Cited on pages 18, 99, and 103.

[KAP05] R. Köhler, G. Altmann, and R.G. Piotrowski, *Quantitative Linguistik. Ein internationales Handbuch / Quantitative Linguistics. An International Handbook*, de Gruyter, Berlin, 2005. Cited on page 11.

[KGW08] J. Ke, T. Gong, and W.S.-Y. Wang, *Language change and social networks*, Commun. Comput. Phys. **3** (2008), 935–949. Cited on page 88.

[Kle47] M. Kleiber, *Body size and metabolic rate*, Physiol. Rev. **27** (1947), 511–541. Cited on page 7.

[Kle13] W. Klein, *Von Reichtum und Armut des deutschen Wortschatzes*, Erster Bericht zur Lage der deutschen Sprache (Deutsche Akademie für Sprache und Dichtung, ed.), de Gruyter, Berlin Boston, 2013. Cited on pages 24, 27, and 112.

[KN11] B. Karrer and M.E.J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E **83** (2011), 016107. Cited on pages 68, 70, and 116.

[KPW07] S. Knapp, A. Polaszek, and M. Watson, *Spreading the word*, Nature **446** (2007), 261–262. Cited on page 106.

[KR13] P.L. Krapivsky and S. Redner, *Distinct degrees and their distribution in complex networks*, J. Stat. Mech. **2013** (2013), P06002. Cited on pages 24 and 115.

[LB14] R. Louf and M. Barthelemy, *Scaling : Lost in the smog*, Environ. and Plann. B **41** (2014), 767–769. Cited on page 8.

[Lev09] M. Levy, *Gibrat's Law for (All) Cities: Comment*, Am. Econ. Rev. **99** (2009), 1672–1675. Cited on page 8.

[LFR08] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Benchmark graphs for testing community detection algorithms*, Phys. Rev. E **78** (2008), 1–5. Cited on page 76.

[Lin91] J. Lin, *Divergence measures based on the Shannon entropy*, IEEE T. Inform. Theory **37** (1991), 145–151. Cited on page 49.

[LMA+12] Y. Lin, J.-B. Michel, E.L. Aiden, J. Orwant, W. Brockman, and S. Petrov, *Syntactic Annotations for the Google Books Ngram Corpus*, Proceedings of the ACL 2012 System Demonstrations - ACL '12 (Stroudsburg, PA), Association for Computational Linguistics, 2012. Cited on pages 87 and 119.

[LMC10] W. Li, P. Miramontes, and G. Cocho, *Fitting Ranked Linguistic Data with Two-Parameter Functions*, Entropy **12** (2010), 1743–1764. Cited on page 16.

[LMJ+07] E. Lieberman, J.-B. Michel, J. Jackson, T. Tang, and M.A. Nowak, *Quantifying the evolutionary dynamics of language*, Nature **449** (2007), 713–6. Cited on pages 85, 86, 102, and 103.

[Lot26] A.J. Lotka, *The frequency distribution of scientific productivity*, J. of Washington Academy Sciences **16** (1926), 317–323. Cited on page 1.

[LPPM11] J. Lijffijt, P. Papapetrou, K. Puolamäki, and H. Mannila, *Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer-Verlag, Berlin Heidelberg, 2011. Cited on pages 11 and 23.

[LSW+15] A. Lancichinetti, M.I. Sirer, J.X. Wang, D. Acuna, K. Körding, and L.A.N. Amaral, *A high-reproducibility and high-accuracy method for automated topic classification*, Phys. Rev. X **5** (2015), 011007. Cited on page 70.

[LZZ10] L. Lü, Z. Zhang, and T. Zhou, *Zipf's law leads to Heaps' law: analyzing their relation in finite-size systems*, PloS one **5** (2010), e14139. Cited on page 24.

[MA14] J.M. Miotto and E.G. Altmann, *Predictability of Extreme Events in Social Media*, PloS one **9** (2014), e111506. Cited on page 8.

[MAAJ13] J. Mathiesen, L. Angheluta, P.T.H. Ahlgren, and M.H. Jensen, *Excitable human dynamics driven by extrinsic events in massive communities*, Proc. Nat. Acad. Sci. USA (2013), 1–4. Cited on page 89.

[Man53] B. Mandelbrot, *An informational theory of the statistical structure of language*, Communication Theory (W. Jackson, ed.), Butterworth, Woburn, MA, 1953. Cited on page 16.

[Man61] ———, *On the theory of word frequencies and on related markovian models of discourse*, Structure of Language and Its Mathematical Aspects: Proceedings of Symposia in Applied Mathematics Vol. XII (R. Jakobson, ed.), American Mathematical Society, Providence, RI, 1961. Cited on pages 21 and 24.

[MBG+94] R. Mantegna, S. Buldyrev, A. Goldberger, S. Havlin, C. Peng, M. Simons, and H. Stanley, *Linguistic Features of Noncoding DNA Sequences*, Phys. Rev. Lett. **73** (1994), 3169–3172. Cited on pages 48 and 116.

[McC02] A. McCallum, *Mallet: A machine learning for language toolkit*, 2002, `http://mallet.cs.umass.edu`. Cited on page 73.

[Mil55] G.A. Miller, *Note on the bias of information estimates*, Information theory in psychology: Problems and methods (H. Quastler, ed.), Free Press, Glencoe, IL, 1955. Cited on pages 52 and 59.

[Mit04] M. Mitzenmacher, *A Brief History of Generative Models for Power Law and Lognormal Distributions*, Internet Math. **1** (2004), 226–251. Cited on pages 7, 8, 21, and 29.

[Mit11] M. Mitchell, *Complexity: A Guided Tour*, Oxford University Press, Oxford, 2011. Cited on page 7.

[MKEHG11] A.P. Masucci, A. Kalampokis, V.M. Eguíluz, and E. Hernández-García, *Wikipedia information flow analysis reveals the scale-free architecture of the semantic space*, PloS one **6** (2011), e17333. Cited on pages 25 and 49.

[MRS08] C.D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, Cambridge, 2008. Cited on pages 2, 24, and 63.

[MS99] C.D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, 1999. Cited on pages 2, 10, 36, 49, and 51.

[MSA+11] J.-B. Michel, Y.K. Shen, A.P. Aiden, A. Veres, M.K. Gray, J.P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M.A. Nowak, and E.L. Aiden, *Quantitative analysis of culture using millions of digitized books*, Science **331** (2011), 176–182. Cited on pages 2, 12, 24, 27, 85, 87, 103, 112, 117, 119, and 122.

[MSFCC15] I. Moreno-Sánchez, F. Font-Clos, and Á. Corral, *Large-scale analysis of Zipf's law in English texts*, arXiv:1509.04486 (2015), 1–11. Cited on page 22.

[MZ10] M.A. Montemurro and D.H. Zanette, *Towards the quantification of the semantic information encoded in written language*, Adv. Complex Syst. **13** (2010), 135. Cited on pages 35, 37, 38, and 113.

[MZL12] S.A. Myers, C. Zhu, and J. Leskovec, *Information diffusion and external influence in networks*, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12 (New York), ACM, 2012. Cited on page 88.

[NB98] S. Naranan and V.K. Balasubrahmanyan, *Models for power law relations in linguistics and information science*, J. Quant. Linguist. **5** (1998), 35–61. Cited on page 20.

[New05] M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemp. Phys. **46** (2005), 323–351. Cited on pages 2, 7, 8, 21, 29, and 58.

[New10] ———, *Networks: An Introduction*, Oxford University Press, Oxford, 2010. Cited on pages 2, 67, 91, and 98.

[New11a] ———, *Communities, modules and large-scale structure in networks*, Nat. Phys. **8** (2011), 25–31. Cited on page 67.

[New11b] ———, *Resource Letter CS–1: Complex Systems*, Am. J. Phys. **79** (2011), 800. Cited on page 7.

[NG04] M.E.J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E **69** (2004), 026113. Cited on page 67.

[Niy06] P. Niyogi, *The Computational Nature of Language Learning and Evolution*, MIT Press, Cambridge, MA, 2006. Cited on page 88.

[Nol15] J.P. Nolan, *Stable distributions - models for heavy tailed data*, Birkhauser, Boston, 2015, In progress, Chapter 1 online at `http://academic2.american.edu/~jpnolan`. Cited on page 9.

[NSB02] I. Nemenman, F. Shafee, and W. Bialek, *Entropy and Inference, Revisited*, Advances in Neural Information Processing Systems 14 (Cambridge, MA), MIT Press, 2002. Cited on page 59.

[OLBC10] F.W.J. Olver, D.W. Lozier, R.F. Boisvert, and C.W. Clark (eds.), *NIST Handbook of Mathematical Functions*, Cambridge University Press, New York, NY, 2010, Print companion to [DLM]. Cited on page 138.

[oPML] University of Pisa Multimedia Lab, *Wikipedia extractor*, `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor` [Online; accessed 28-February-2013]. Cited on page 119.

[PAM07] M. Pagel, Q.D. Atkinson, and A. Meade, *Frequency of word-use predicts rates of lexical evolution throughout Indo-European history*, Nature **449** (2007), 717–20. Cited on pages 85 and 86.

[PAOP10] M. Prokopenko, N. Ay, O. Obst, and D. Polani, *Phase transitions in least-effort communications*, J. Stat. Mech. **2010** (2010), P11025. Cited on page 3.

[Par96] V. Pareto, *Cours d'Economie Politique*, Droz, Geneva, 1896. Cited on pages 1 and 7.

[PDD15a] E.A. Pechenick, C.M. Danforth, and P.S. Dodds, *Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution*, PloS one **10** (2015), e0137041. Cited on page 13.

[PDD15b] ———, *Is language evolution grinding to a halt?: Exploring the life and death of words in English fiction*, arXiv:1503.03512 (2015). Cited on page 49.

[Pei14a] T.P. Peixoto, *The graph-tool python library*, figshare (2014), `http://figshare.com/articles/graph_tool/1164194`. Cited on page 74.

[Pei14b] ———, *Hierarchical Block Structures and High-Resolution Model Selection in Large Networks*, Phys. Rev. X **4** (2014), 011047. Cited on pages 68 and 69.

[Pei15] ———, *Model selection and hypothesis testing for large-scale network models with overlapping groups*, Phys. Rev. X **5** (2015). Cited on pages 68 and 80.

[Pia14] S.T. Piantadosi, *Zipf's word frequency law in natural language: A critical review and future directions*, Psychon. B. Rev. **21** (2014), 1112–30. Cited on pages 3 and 11.

[Pin99] S. Pinker, *Words and Rules: The Ingredients of Language*, Basic Books, New York, 1999. Cited on pages 102 and 122.

[PJSB13] J.I. Perotti, H.H. Jo, A.L. Schaigorodsky, and O.V. Billoni, *Innovation and nested preferential growth in chess playing behavior*, Europhys. Lett. **104** (2013), 48005. Cited on page 24.

[PSCvV15] R. Pastor-Satorras, C. Castellano, P. van Mieghem, and A. Vespignani, *Epidemic processes in complex networks*, Rev. Mod. Phys. **87** (2015), 925–979. Cited on page 114.

[PSD14] J.B. Pierrehumbert, F. Stonedahl, and R. Daland, *A model of grassroots changes in linguistic systems*, arXiv:1408.1985 (2014). Cited on page 88.

[PTG11] S.T. Piantadosi, H. Tily, and E. Gibson, *Word lengths are optimized for efficient communication*, Proc. Nat. Acad. Sci. USA **108** (2011), 3526–9. Cited on page 11.

[PTH+12] A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, and M. Perc, *Languages cool as they expand: allometric scaling and the decreasing need for new words*, Sci. Rep. **2** (2012), 943. Cited on pages 20, 24, and 29.

[Rén61] A. Rényi, *On Measures of Entropy and Information*, Berkeley Symposium on Mathematical Statistics and Probability (Berkeley, CA), University of California Press, 1961. Cited on page 50.

[Rog03] E.M. Rogers, *Diffusion of Innovations*, 5th ed., Free Press, New York, 2003. Cited on pages 88 and 106.

[ŘS10] R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora*, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (Valletta, Malta), ELRA, May 2010, `http://is.muni.cz/publication/884893/en`, pp. 45–50 (English). Cited on pages 38 and 43.

[RSOH12] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf, *Text-mining solutions for biomedical research: enabling integrative biology*, Nat. Rev. Genet. **13** (2012), 829–839. Cited on page 63.

[SB04] R.M. Shiffrin and K. Börner, *Mapping knowledge domains*, Proc. Nat. Acad. Sci. USA **101 Suppl** (2004), 5183–5. Cited on pages 2 and 63.

[SCBn+12] J. Serrà, Á. Corral, M. Boguñá, M. Haro, and J.L. Arcos, *Measuring the Evolution of Contemporary Western Popular Music*, Sci. Rep. **2** (2012), 1–6. Cited on pages 48 and 116.

[Sch78] G. Schwarz, *Estimating the Dimension of a Model*, Ann. Stat. **6** (1978), 461–464. Cited on page 95.

[Sch04] T. Schürmann, *Bias Analysis in Entropy Estimation*, J. Phys. A: Math. Gen. (2004), 5. Cited on page 52.

[SCMF10] R.V. Solé, B. Corominas-Murtra, and J. Fortuny, *Diversity, competition, extinction: the ecophysics of language change*, J. R. Soc. Interface **7** (2010), 1647–64. Cited on page 88.

[SCMVS09] R.V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, *Language Networks: their structure, function and evolution*, Complexity **15** (2009), 20–26. Cited on page 11.

[SDGA04] D. Sornette, F. Deschâtres, T. Gilbert, and Y. Ageon, *Endogenous Versus Exogenous Shocks in Complex Networks: An Empirical Test Using Book Sale Rankings*, Phys. Rev. Lett. **93** (2004), 1–4. Cited on page 89.

[SDM01] J.P. Sethna, K.A. Dahmen, and C.R. Myers, *Crackling noise*, Nature **410** (2001), 242–50. Cited on pages 7 and 8.

[SFM09] M.A. Serrano, A. Flammini, and F. Menczer, *Modeling statistical properties of written text*, PloS one **4** (2009), e5372. Cited on pages 24 and 25.

[Sha48] C.E. Shannon, *A Mathematical Theory of Communication*, Bell Syst. Tech. J. **27** (1948), 379–423,623–656. Cited on page 12.

[Sim55] H.A. Simon, *On a class of skew distribution functions*, Biometrika **42** (1955), 425–440. Cited on page 25.

[Sor06] D. Sornette, *Critical Phenomena in Natural Sciences*, Springer-Verlag, Berlin Heidelberg, 2006. Cited on pages 2, 7, 8, and 39.

[SP12] M.P.H. Stumpf and M.A. Porter, *Critical Truths About Power Laws*, Science **335** (2012), 665–666. Cited on pages 2, 8, 23, 113, and 114.

[SR10] M.V. Simkin and V.P. Roychowdhury, *Re-inventing Willis*, Phys. Rep. **502** (2010), 1–35. Cited on page 29.

[Sta99] H. Stanley, *Scaling, universality, and renormalization: Three pillars of modern critical phenomena*, Rev. Mod. Phys. **71** (1999), S358–S366. Cited on pages 2 and 7.

[SZZ93] A. Schenkel, J.U.N. Zhang, and Y.-C. Zhang, *Long range correlation in human writings*, Fractals **1** (1993), 47. Cited on pages 11 and 23.

[Tay61] L.R. Taylor, *Aggregation, Variance and the Mean*, Nature **189** (1961), 732–735. Cited on page 38.

[Tay97] J.R. Taylor, *An Introduction to Error Analysis*, University Science Books, Sausalito, CA, 1997. Cited on page 19.

[TLSS14] F. Tria, V. Loreto, V.D.P. Servedio, and S.H. Strogatz, *The dynamics of correlated novelties*, Sci. Rep. **4** (2014), 1–44. Cited on pages 24 and 115.

[Tsa88] C. Tsallis, *Possible generalization of Boltzmann-Gibbs statistics*, J. Stat. Phys. **52** (1988), 479–487. Cited on page 50.

[Tul96] J. Tuldava, *The frequency spectrum of text and vocabulary*, J. Quant. Linguist. **3** (1996), 38–50. Cited on page 16.

[VA12] N.K. Vitanov and M.R. Ausloos, *Knowledge Epidemics and Population Dynamics Models for Describing Idea Diffusion*, Models of Science Dynamics (A. Scharnhorst, K. Börner, and P. Besselaar, eds.), Springer-Verlag, Berlin Heidelberg, 2012. Cited on pages 88 and 106.

[vv05] D.C. van Leijenhorst and T.P. van der Weide, *A formal derivation of Heaps' Law*, Inform. Sciences **170** (2005), 263–272. Cited on pages 24 and 25.

[WA99] G. Wimmer and G. Altmann, *Review Article: On Vocabulary Richness*, J. Quant. Linguist. **6** (1999), 1–9. Cited on pages 24 and 45.

[Wei78] M.S. Weiss, *Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data*, J. Am. Stat. Assoc. **73** (1978), 872–875. Cited on page 23.

[Wes97] G.B. West, *A General Model for the Origin of Allometric Scaling Laws in Biology*, Science **276** (1997), 122–126. Cited on page 7.

[Wik] Wikimedia, *Dump of the english wikipedia on 02-june-2012*, `http://dumps.wikimedia.org/enwiki/` [Online; accessed 26-February-2013]. Cited on pages 13, 119, and 125.

[Wik14a] Wikipedia, *Constrained writing — Wikipedia, the free encyclopedia*, 2014, [Online; accessed 03-December-2014]. Cited on page 11.

[Wik14b] ———, *German orthography reform of 1996 — Wikipedia, the free encyclopedia*, 2014, [Online; accessed 13-June-2014]. Cited on pages 101 and 102.

[Wik14c] ———, *Romanization of russian — Wikipedia, the free encyclopedia*, 2014, [Online; accessed 13-June-2014]. Cited on page 102.

[WLH68] U. Weinreich, W. Labov, and M.R. Herzog, *Empirical Foundations for a Theory of Language Change*, Directions for Historical Linguistics (W. Lehmann and Y. Malkiel, eds.), The University of Texas Press, Austin, TX, 1968. Cited on pages 3, 83, 87, and 89.

[WMM09] H.M. Wallach, D. Mimno, and A. Mccallum, *Rethinking LDA : Why Priors Matter*, Advances in Neural Information Processing Systems 22 (NIPS 2009), 2009. Cited on pages 66 and 78.

[WPC+15] N.W. Watkins, G. Pruessner, S.C. Chapman, N.B. Crosby, and H.J. Jensen, *25 Years of Self-organized Criticality: Concepts and Controversies*, Space Sci. Rev. (2015), 1–42. Cited on page 8.

[WPCG+14] C.H. Weiss, J. Poncela-Casasnovas, J.I. Glaser, A.R. Pah, S.D. Persell, D.W. Baker, R.G. Wunderink, and L.A.N. Amaral, *Adoption of a High-Impact Innovation in a Homogeneous Population*, Phys. Rev. X **4** (2014), 041008. Cited on page 115.

[WZ05] H.E. Williams and J. Zobel, *Searchable words on the Web*, Int. J. Digit. Libr. **5** (2005), 99–105. Cited on page 24.

[YKK12] T. Yasseri, A. Kornai, and J. Kertész, *A Practical Approach to Language Complexity: A Wikipedia Case Study*, PLoS ONE **7** (2012), e48386. Cited on pages 24 and 45.

[Yul25] G.U. Yule, *A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S.*, Philos. T. R. Soc. B **213** (1925), 21–87. Cited on page 29.

[Zan14] D.H. Zanette, *Statistical Patterns in Written Language*, arXiv:1412.3336 (2014). Cited on pages 11 and 38.

[Zip36] G.K. Zipf, *The Psycho-Biology of Language*, Routledge, London, 1936. Cited on pages 1, 11, 15, and 29.

[Zip49] ———, *Human Behaviour and the Principle of Least Effort*, Addison-Wesley, Reading, MA, 1949. Cited on pages 3, 7, 11, and 15.

[ZM05] D.H. Zanette and M. Montemurro, *Dynamics of Text Generation with Realistic Zipf's Distribution*, J. Quant. Linguist. **12** (2005), 29–40. Cited on pages 24, 25, and 31.

[ZMN15] P. Zhang, C. Moore, and M.E.J. Newman, *Community detection in networks with unequal groups*, arXiv:1509.00107 (2015). Cited on page 114.

Acknowledgements

Foremost, I am indebted to Eduardo G. Altmann whom I cannot thank too much for his inspiring supervision of this thesis; besides the support and guidance I appreciate the way we could develop and discuss new ideas and continuously challenge our understanding of old ones.

Likewise, I would like to thank my collaborators, Francesc Font-Clos, Fakhteh Ghanbarnejad, Jose M. Miotto, and Tiago P. Peixoto, who were directly involved in projects that led to results presented in this thesis.

Further, I had the chance to interact with many people by sharing views, discussing ideas, or learning from each other - thanks to everyone; this includes the members of the group Dynamical Systems & Social Dynamics, everyone involved in the Complex Systems Reading Group at the MPIPKS, as well as Joachim Scharloth and Noah Bubenhofer.

I also want to thank everyone, colleagues and staff alike, at the Max Planck Institute for the Physics of Complex Systems for contributing to such a pleasant and stimulating environment.

Finally, I want to thank Antje for everything.

Versicherung

Diese Arbeit wurde am Max-Planck-Institut für Physik komplexer Systeme unter der wissenschaftlichen Betreuung von Prof. Dr. Jan-Michael Rost durchgeführt.
Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.
Darüber hinaus erkenne ich die Promotionsordnung der Fakultät Mathematik und Naturwissenschaften der Technischen Universität Dresden vom 23. Februar 2011 an.

Martin Gerlach

Dresden, 22. Oktober 2015