BOSTON UNIVERSITY
GRADUATE SCHOOL OF ARTS AND SCIENCES

Dissertation

# INTERDISCIPLINARY APPLICATIONS OF STATISTICAL PHYSICS TO COMPLEX SYSTEMS: SEISMIC PHYSICS, ECONOPHYSICS, AND SOCIOPHYSICS

by

**JOEL TENENBAUM**

B.A., Goucher College, 2006
M.A., Boston University, 2008

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
2012

Approved by

First Reader _____

H. Eugene Stanley, Ph.D.
University Professor and Professor of Physics

Second Reader _____

William Skocpol Ph.D.
Professor of Physics

# Acknowledgments

I would to like to acknowledge my parents, Arthur and Judie, for raising me, for bearing with me during the challenges, and for joining me in the triumphs, for letting me be a bit too ridiculous occasionally when my passions led me there, for never damping my sense of curiosity, for never imposing, for never having anything but confidence in me. I thank my mom for the endless homebaked cookies and for showing me the importance of valuing the people in my life, my dad for infecting me with a critical eye and the need to know how things and humans work. I dedicate this thesis to you both.

I would like to thank my thesis advisor, Professor Gene Stanley for taking me into his family of impassioned collaborators, for endless champagne celebrations of life and endless pizza celebrations of science, for teaching by example how to value people, for passionately being his genuine self, for the sense of family that this engendered in his group, for his indefatigable work ethic and boundless curiosity and enthusiasm, for his over-lavish encouragement and thoughtfully over-delicate feedback.

I acknowledge my collaborators, Shlomo Havlin and Boris Podobnik, who give so freely of their time and their talents. Thanks to Bob Tomposki and Jerry Morrow for their crucial help and wonderful presence. Thank you so much to Mirtha Cabello for tirelessly picking up the pieces. Without you, everything would collapse.

Thank you to my dissertation committee, Professors Sheldon Glashow, Rick Averitt, Anders Sanvik, Gene Stanley, and Bill Skocpol, for their patience. An additional thank you to thesis readers Bill Skocpol and Gene Stanley.

I thank my undergraduate advisor, Sasha Dukan, for her words of encouragement and enthusiasm for physics, and for her wise suggestion that I apply to Boston

University. You were right: these have been the happiest years of my life.

Many thanks to Charlie Nesson for his daring and wizardous rescue, for his encouragement to me for my scientific career, and for his unconditional confidence in me. You are Atticus Finch. You are Vincent Laguardia Gambini. You are Gandalf. Thanks to Debbie Rosenbaum, Fern, Jason Harrow, Isaac Meister, Phil Hill, and all the others who've helped defend me over the years, while studying what you love so that I can study what I love. I was glad to be your homework.

To the amazing friends who have spent this period of my life with me who made Boston my home. On every street are the countless memories of the past six years. To the good friends we've had and good friends we've lost along the way: Alex, both the voice of reason when it was needed and the voice of levity when it was needed, I'll miss Jersey Shore episodes and George Foreman-cooked Trader Joes meat. Mason, my co-conspirator in both my endless appetite for erudite conversation and skepticism that grounds my sanity. Erik, your patience and assistance in all things coding saved me many times. Ashli, Maria, Ying, Annie, Erin, Jiayuan, Sean, Matt, Jordan, Erol, Diego, Elena, Heather, Ashley, Javad, Nagendra, Felipe, Deepani, Vidya, Mark, Meghan, Guci, Ashleigh, Naomi, Dan, Amy, I've always felt like no one but myself around you guys and that's really meant a lot to me. Thank you to Shambhala and the Shambhala community. Kelly and Alvaro, I will always remember you. You were some of my favorite people in the world.

I would like to acknowledge my sister Abi, for valuable advice in college. I'm happy for you and Gabe, and for your and Gabe's recent priceless additions: Louisa Antoinette and Benjamin Sullivan. They have a good mother and a caring father. I would also like to acknowledge my sister Tova who somehow so thoroughly forgave me for years of both being an older brother and being a physicist.

# APPLICATIONS OF STATISTICAL PHYSICS TO COMPLEX SYSTEMS: SEISMIC PHYSICS, ECONOPHYSICS, AND SOCIOPHYSICS

(Order No. )

**JOEL TENENBAUM**

Boston University Graduate School of Arts and Sciences, 2012

Major Professor: H. Eugene Stanley, University Professor and Professor of Physics

## ABSTRACT

This thesis applies statistical physics concepts and methods to quantitatively analyze complex systems.

This thesis is separated into four parts: (i) characteristics of earthquake systems (ii) memory and volatility in data time series (iii) the application of part (ii) to world financial markets, and (iv) statistical observations on the evolution of word usage.

In Part I, we observe statistical patterns in the occurrence of earthquakes. We select a 14-year earthquake catalog covering the archipelago of Japan. We find that regions traditionally thought of as being too distant from one another for causal contact display remarkably high correlations, and the networks that result have a tendency to link highly connected areas with other highly connected areas.

In Part II, we introduce and apply the concept of "volatility asymmetry", the primary use of which is in financial data. We explain the relation between memory and "volatility asymmetry" in terms of an asymmetry parameter $\lambda$. We define a litmus test for determining whether $\lambda$ is statistically significant and propose a stochastic model based on this parameter and use the model to further explain empirical data.

In Part III, we expand on volatility asymmetry. Importing the concepts of time dependence and universality from physics, we explore the aspects of emerging (or "transition") economies in Eastern Europe as they relate to asymmetry. We find that these emerging markets in some instances behave like developed markets and in other instances do not, and that the distinction is a matter both of country and a matter of time period, crisis periods showing different asymmetry characteristics than "healthy" periods.

In Part IV, we take note of a series of findings in econophysics, showing statistical growth similarities between a variety of different areas that all have in common the fact of taking place in areas that are both (i) competing and (ii) dynamic. We show that this same growth distribution can be reproduced in observing the growth rates of the usage of individual words, that just as companies compete for sales in a zero sum marketing game, so do words compete for usage within a limited amount of reader man-hours.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| ARCH | AutoRegressive Conditional Heteroskedasticity |
| DFA | detrended fluctuation analysis |
| EEG | electroencephalography |
| GARCH | Generalized AutoRegressive Conditional Heteroskedasticity |
| i.i.d. | independent and identically distributed |
| JUNEC | Japan University Network Earthquake Catalog |
| MLE | maximum likelihood estimation |
| pdf | probability density function |

# Part I

# Introduction

# Chapter 1

# Interdisciplinary Applications of Statistical Physics to Complex Systems: Seismic Physics, Econophysics, and Sociophysics

Interdisciplinary science is an emerging field which applies methods from various fields to address research topics that have traditionally been analyzed by scientists within a specific field. Specifically, within interdisciplinary research and growing in import is the concept of complexity. While various definitions of complexity have been expounded most have in common the notion of a large number of distinct parts interacting at different levels within the framework of a hierarchy that operates on a wide gamut of scales. Often these interactions are fractal in nature, manifesting in the form of power-law relationships between the variables under consideration. This stands in stark contrast to noncomplex systems, such as the motion of a random walker or the kinematic statistics of an ideal gas which, though they may be complicated, are characterized by uniscaled functional forms like exponentials. The power-law form is central to complexity because it displays a scale invariance not found in exponential or related forms. Changing the input scale by a factor of $\lambda$, for example results in qualitatively identical behavior, differing only by a constant factor:

$$f(\lambda x) = A(\lambda x)^k = \lambda^k f(x) \propto f(x). \tag{1.1}$$

As a consequence, the relation obeys a scaling law in $\lambda$. $\lambda$ can be "scaled out" in that $\frac{f(\lambda x)}{\lambda^k}$ has no dependence on $\lambda$. If, for example, f is a relationship that tells a company's growth rate, based on its size $x$, and $\lambda$ can be scaled out in this manner, it shows that companies of all sizes obey the same underlying mechanics.

The most obvious philosophical consequence of scale invariance is the lack of a characteristic size for observations. While an exponential, for example, has a characteristic scale, at which one may consider further effects to be nominal (e.g. the half-life of a radioactive particle, the penetration depth of electromagnetic radiation into a material), a power-law exhibits identical qualitative behavior at any scale, resulting in relationships that do not vary over several orders of magnitude.

Methods from statistical physics are well-suited for interdisciplinary complexity research due to the similarity between physical systems consisting of interacting particles and complex systems consisting of interacting constituents. In this work, we focus our attention on three such systems that display emergent complexity arising from interactions between the constituent parts:

(i) the emergent behavior in seismology resulting from the intricate distribution of stress along the world's tectonic plates and associated interplate and intraplate fault systems,

(ii) the interaction between millions of interacting individuals manifesting in stock market prices,

(iii) the collective dynamics resulting from words in a language competing for usage.

We use statistical physics concepts such as scaling analysis, universality, symmetry, stationarity, and networks, as well as tools from econometrics such as cross-correlations, autoregressive hetereoskedastic processes to analyze these three systems. The purpose of this research is to demonstrate the utility of these methods in explaining data recorded for various complex systems and to find statistical laws that quantify the statistical regularities we observe.

Our research is a combination of both theoretical modeling and experimental observation. We utilize comprehensive data recorded for

(i) seismic records of earthquake events over the 14-year period 1985-1998 as recorded by nine institutions of higher learning across Japan,

(ii) the financial time series of stock market indices for 11 Eastern European transition economies over the 20-year period 1989-2009, as well as 3 United States market indices for the 27-year period 1980-2008

(iii) word instances of $1 \times 10^7$ distinct words from $10^6$ digitized books, over the 209-year period 1800-2008

and find several statistical laws. These empirical descriptions aid in the development of appropriate theoretical models, which can provide further insight into the various deterministic and stochastic mechanisms that give rise to real-world phenomena. We also invoke electroencephalography (EEG) data for 25 subjects during an overnight hospital stay as a "proof of concept" for a statistical model we introduce.

# Part II

# Earthquake networks based on similar activity patterns

# Chapter 2

# Background

One of earliest observed phenomena to exhibit what we now call "complexity" is that of earthquakes. While many natural phenomena known to antiquity occurred with clockwork regularity such as tides, seasons, or even the flooding of the Nile Delta, the pattern behind earthquake behavior seemed to have neither regularity or even an approximate time scale upon which to plan or make any kind of headway into understanding. Consequently, many ancient peoples posited explanations involving large animals supporting the world or instead explained earthquakes as the capricious kicks of a large fetus in an enormous womb. These explanations, animistic at their root, all tacitly reinforce the perception of earthquakes as arbitrary and unpredictable events.

Since the advent of modern study, progress has been made. Despite the underlying complexities of earthquake dynamics, celebrated statistical scaling laws have emerged, describing the number of events of a given magnitude (Gutenberg-Richter law)[1], the decaying rate of aftershocks after a main event (Omori law)[2], the magnitude difference between the main shock and its largest aftershock (Bath law)[3], as well as the fractal spatial occurrence of events[4, 5]. Indeed, Bak et al. have unified these three of these laws[6, 7] by scaling out spatiotemporal information, revealing an underlying common structure. However, while the fractal occurrence of earthquakes incorporates spatial dependence, it ostensibly embeds isotropy in the form of radial symmetry, while real-world earthquakes are usually anisotropic[8].

To better characterize this anisotropic spatial dependence as it applies to such

heterogeneous geography, network approaches have been recently applied to study earthquake catalogs[9, 10, 11, 12, 13]. These recent network approaches define links as being between successive events or being between events which have a relatively small probability of both occurring based on three of the above statistical scaling laws[14]. These methods define links between singular events. In contrast, we define links between locations based on long-term similarity of earthquake activity. While earlier approaches capture the dynamic nature of an earthquake network, they do not incorporate the characteristic properties of each particular location along the fault[15]. Various studies have shown[15, 16, 17, 18, 2] that localized areas within a catalog have non-Poissonian clustering in time, even within aftershock sequences[18], demonstrating that each area not only has its own statistical characteristics[20], but also retains a memory of its events[15, 16, 17]. As a result, successive events may not be just the result of uncorrelated independent chance but instead might be dependent on the history particular to that location. If prediction is to be a goal of earthquake research, it makes sense to incorporate such long-term behavior inherent to a given location by integrating the analysis of each location over time, rather than by treating each event independently. We include long-term behavior as such in this paper by considering a network of locations, where each location is characterized by its long-term activity over several years.

# Chapter 3

# Data

For our analysis, we utilize data from the Japan University Network Earthquake Catalog (JUNEC), available online at
http://wwweic.eri.u-tokyo.ac.jp/CATALOG/junec/. We choose the JUNEC catalog because Japan is among the most active and best observed seismic regions in the world. Because our technique is novel, this catalog provided the best first avenue to analysis. In the future, it may be possible to fine-tune our approach to more sparse catalogs.

The data in the JUNEC catalog span 14 years from 1 July 1985 - 31 December 1998. Each line of the catalog includes the date, time, magnitude, latitude, and longitude of the event. We found the catalog to obey the Gutenberg-Richter law for events of magnitude 2.2 or larger. By convention, this is taken to mean that the catalog can be assumed to be complete in that magnitude range. However, because catalog completeness cannot be guaranteed for shorter time periods over a 14-year span, we also examine Gutenberg-Richter statistics for each nonoverlapping two-year period (Fig 3.1). We find that, though absolute activity varies by year, the relative occurrences of quakes of varying magnitudes does not change significantly for events between magnitude 2.2 and 5, where there is the greatest danger of events missing from the catalog. We also note the spatial clustering of data in the regime of $2.2 \leq M \leq 2.5$. We therefore compare our results to results obtained using 2.5 as a lower bound for event inclusion. See further discussion below.

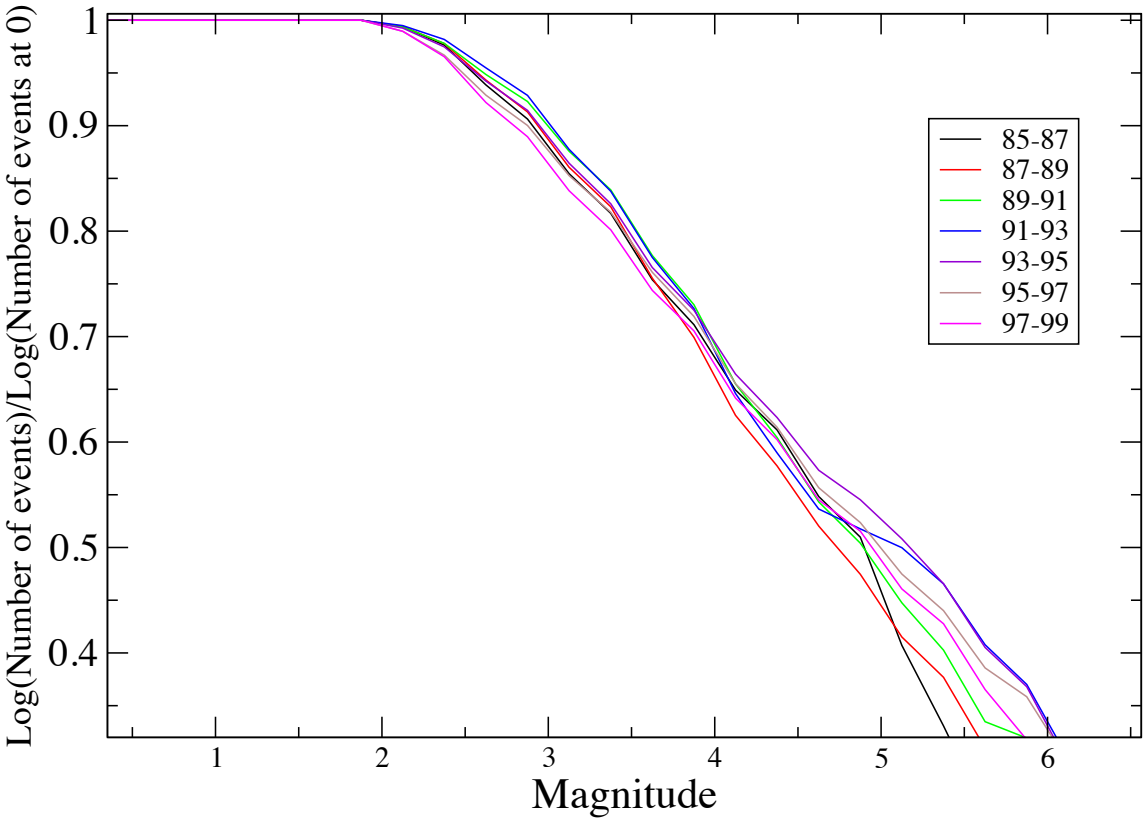


**Figure 3.1**: (Color online) Gutenberg-Richter statistics for the JUNEC catalog every 2 years demonstrating that the magnitude above which the Gutenberg-Richter law is obeyed is approximately constant from year to year.

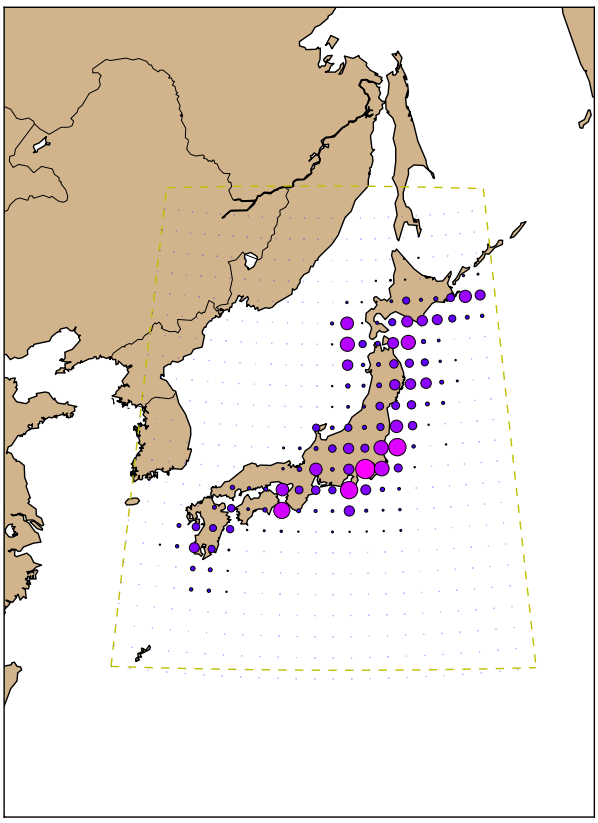

**Figure 3.2**: (Color online) Number of events by location in the JUNEC catalog. The JUNEC catalog clusters spatially, with most activity occurring on the eastern side of Honshu, Japan's principle island.

# Chapter 4

# Method

We partition the region associated with the JUNEC catalog as follows: we take the northernmost, southernmost, easternmost, and westernmost extrema of events in the catalog as the spatial bounds for our analysis. We partition this region into a $23 \times 23$ grid which is evenly spaced in geographic coordinates. Each grid square of approximate size 100 km $\times$ 100 km is regarded as a possible node in our network. Results do not qualitatively differ when the fineness of the spatial grid is modified.

For a given measurement at time $t$, an event of magnitude $M$ occurs inside a given grid square. Similar to the method of Corral[18], we define the signal of a given grid square to form a time series $\{s_t\}$ where each series term $s_t$ is related to the earthquake activity that takes place inside that grid square within the time period $t$.

Because events do not generally occur on a daily basis in a given grid square, it is necessary to bin the data to some level of coarseness. How coarse the data are treated involves a trade-off between precision and data richness.

We obtain the best results, corresponding to the most prominent cross-correlations, by choosing 90 days as the coarseness for our time series. This choice means that $s_t$ will cover a time period of 90 days and $s_{t+1}$ will cover the 90-day non-intersecting time period immediately following, giving approximately 4 increments per year. Results do not qualitatively differ by changing the time coarseness.

Accordingly, we develop a definition of the signal for the time series $\{s_t\}$ belonging to each grid cell $ij$:

"energy released":

$s_t(ij) \equiv \sum_{l=1}^{N_t(ij)} 10^{\frac{3}{2}M_t^l(ij)}$. We choose this definition because the term $10^{\frac{3}{2}M}$ is proportional to the energy released from an earthquake of magnitude M.

We also instituted our analysis with another three signal definitions that we omit here:

(a) "average magnitude":

$s_t(ij) \equiv \frac{1}{N_{ij}} \sum_{\ell=1}^{N_{ij}} M_t^\ell(ij)$ where $M_t^\ell(ij)$ is the magnitude of the event and $N_t(ij)$ is the total number of events occurring in the 90-day time window $t$ in the grid square $ij$.

(b) "number of events":

$s_t(ij) \equiv N_t(ij)$ with the symbols as defined in (a).

(c) "magnitude sum":

$s_t(ij) \equiv \sum_{l=1}^{N_t(ij)} M_t^l(ij)$.

All three of these alternative definitions failed to give results significantly better than the shuffled data that were robust in the various adjustable parameters.

To define a link between two grid squares, we calculate the Pearson product-moment correlation coefficient $r$ between the two time series associated with those grid squares:

$$r_{x,y} \equiv \sum_i \frac{(x_i - \mu_x)(x_i - \mu_y)}{\sigma_x \sigma_y} \tag{4.1}$$

where $\mu_x, \mu_y$ are the means and $\sigma_x, \sigma_y$ are the standard deviations of the series $x, y$.

We consider the two grid squares linked if this cross-correlation is larger than a specified threshold value $r_c$, where $r_c$ is a tunable parameter. As is standard in network-related analysis, we define the degree $k$ of a node to be the number of links the node has. Note that because our signal definition involves an exponentiation of numbers of order 1, energy released, and therefore the cross-correlation between two signals is dominated by large events. Examples of signals with high correlation are shown in Fig 5.1.

To confirm the statistical significance of $r_{x,y}$, we compare $r_{x,y}$ of any two given signals with $r_{x,y}$ calculated by shuffling the time orders of one of the signals. We also compare $r_{x,y}$ with the cross-correlation $r_{x,y}(\tau)$ we obtain by time-shifting one of the signals by varying time increments $\tau$, $r_{x,y}(\tau) \equiv r(s_{1,t}, s_{2,t+\tau})$, where we impose periodic boundaries $t + \tau \equiv (t + \tau) \mod t_{max}$ where $t_{max}$ is the length of the series.

We find that either shifting or shuffling the signal reduces cross-correlation to very low levels (Fig. 5.2). Over the 14-year time period 1985-1998, the overall observed activity increased in the areas covered by the catalog. To ensure that the cross-correlations we calculate are not simply the result of trends in the data, we compare our results to those obtained with linearly detrended data [19]. We find that the trends do not have a significant effect. For example, using $r_c = 0.7$, we obtain 815 links, while detrending the data results in only 3 links dropping below the threshold correlation value. For $r_c = 0.6$, we obtain 1003 links, while detrending results in only 3 links dropped. Additionally after detrending, 94% of correlation values stay within 2% of their values.

# Chapter 5

# Results

As described above, we compare $r_{x,y} \equiv r_{x,y}(0)$ between signals at different locations at the same point in time with $r_{x,y}(\tau)$. Even time-shifting by a single time step (representing 90 days) reduces the cross-correlation to within the margin of significance, as shown in Fig. 5.2. We also find a large number of links with cross-correlations far higher than their shuffled counterparts. The number of links exceed those of time-shuffled data by roughly 2-4 $\sigma$, depending on choice of $r_c$ as shown in Fig. 5.3.

As can be seen in Fig. 5.4, a significant fraction of these links connect nodes further than 1000 km apart, which is consistent with the finding that there is no characteristic cut-off length for interactions between events [14]. This is corroborated by Fig. 5.6, showing the number of links a network has at a given distance as a fraction of the number of links that are geometrically possible. Distances shorter than 100km have sparse statistics due to the coarseness of the grid while distances greater than 2300km have sparse statistics due to the finite spatial extent of the catalog.

Our results, shown in Fig. 5.4, are anisotropic, with the majority of links occurring at approximately 37.5 degrees east of north, which is roughly along the principal axis of Honshu, Japan's main island, and parallel to the highly active fault zone formed by the subduction of the Pacific and Philippine tectonic plates under the Amurian and Okhotsk plates. High degree nodes (i.e. nodes with a large number of links) tend to be found in the northeast and north-central regions of the catalog.

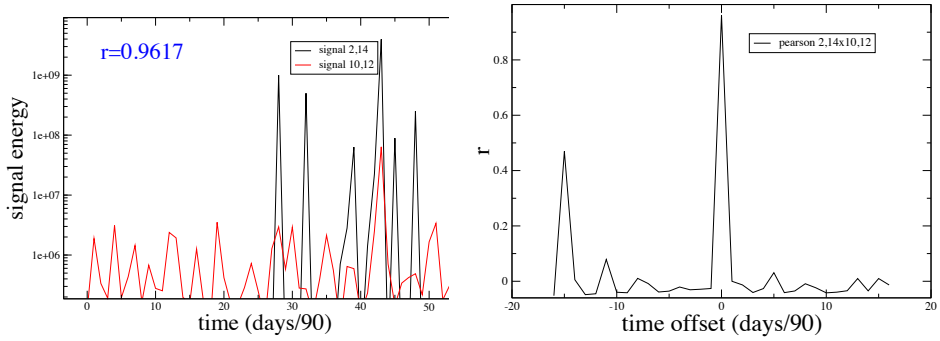


**Figure 5.1**: (Color online) Examples of highly correlated signals, values of (i,j) marked above: (a) Two signals with correlation r=0.9617, 878km apart, (b) corresponding Pearson correlation as a function of time offset. Note that because the signal is defined in terms of exponentiation that large events dominate the correlation, just as large events dominate the total energy released.

Additionally, in network science we often characterize networks by the preference for high-degree nodes to connect to other high-degree nodes. The strength of this preference is quantified by the network's assortativity, with assortativity being defined as:

$$A \equiv r_{k_1,k_2} \tag{5.1}$$

where each link $i$ in the network is described as a link between a node of degree $k_1(i)$ and $k_2(i)$, and r is the Pearson product-moment correlation coefficient as defined in Equation 4.1. Hence, if each node of degree k connects only to nodes of the same degree, the two series $k_1$ and $k_2$ will be identical and A=1.

As shown in Fig 5.5, the networks that result from our procedure are highly assortative with assortativity generally increasing with $r_c$. For comparison we show the assortativity obtained by using the time shuffled networks. Since assortativity of the original networks is far higher than those of shuffled systems, the high assortativity cannot be due to a finite size effect or the spatial clustering displayed in the data (time shuffling preserves location).

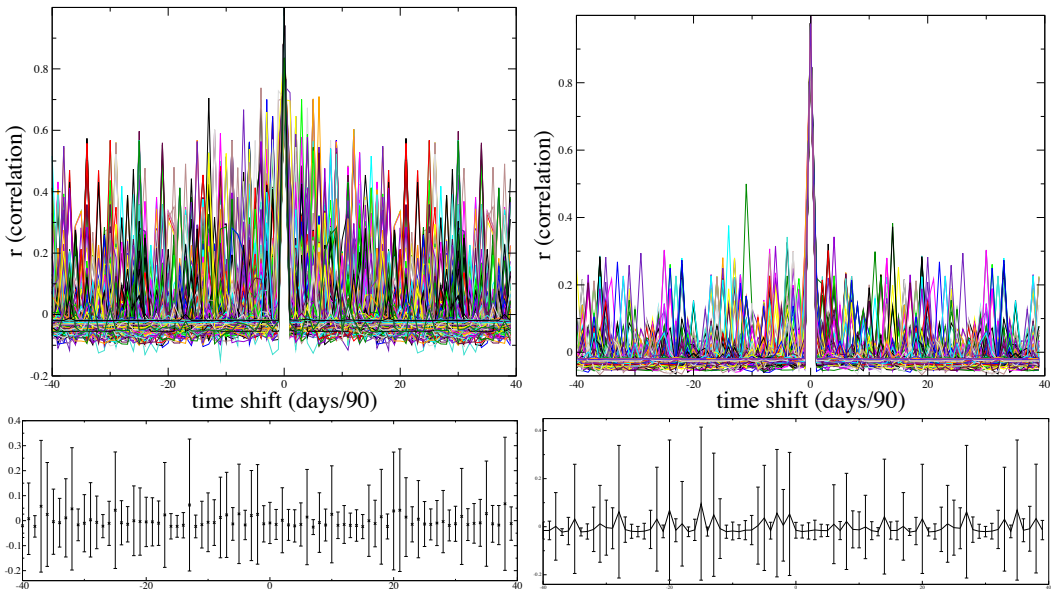


**Figure 5.2**: (Color online) Testing the statistical significance of cross-correlations. For each pair of signals with a correlation larger than $r_c$, we shift one of the signals in time and calculate the new correlation. Offsetting the signals in time results in lower cross-correlation, dropping to the level of noise in the actual data. As a control, we shuffle the signals and calculate the cross-correlation for different time shifts (shown below each figure). Cross-correlation between various pairs of signals vs. time offset. Shown are links for which (a) $r(0) \geq r_c = 0.7$ and (b) $r(0) \geq r_c = 0.9$.

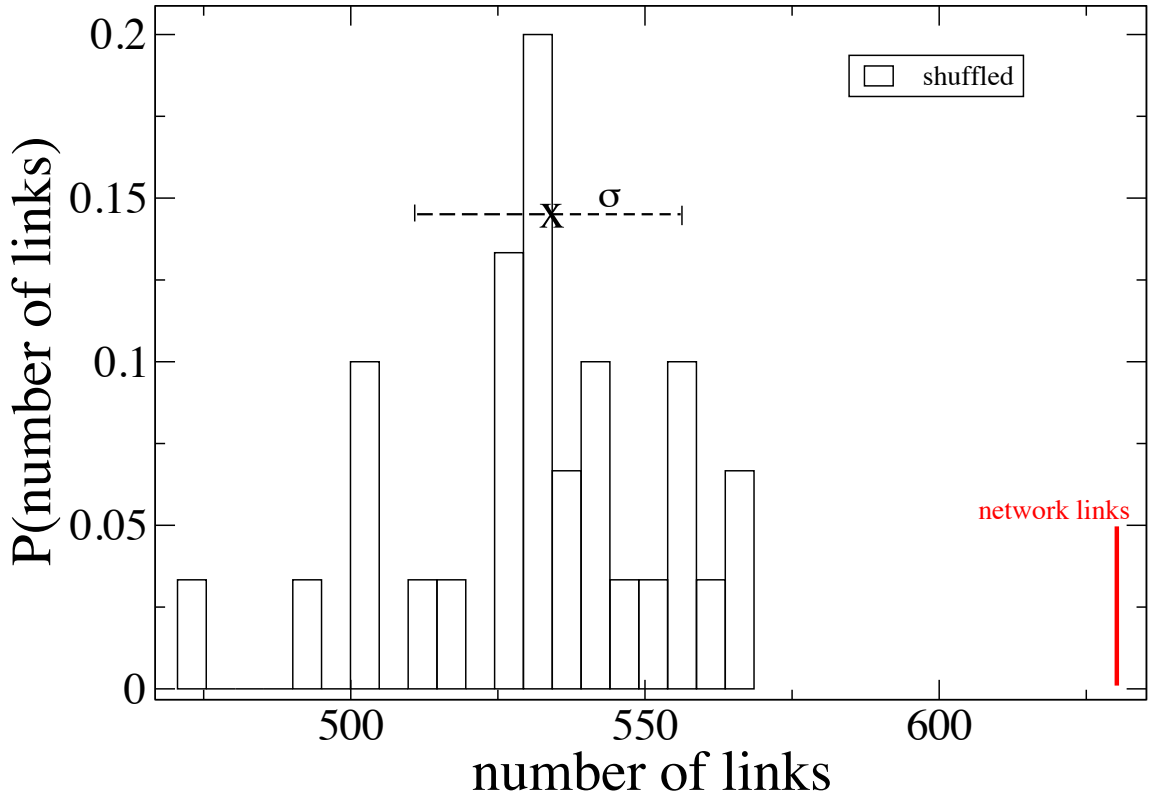


**Figure 5.3**: (Color online) Demonstration that empirical data show far more links than time shuffled data. In black is the distribution of the number of links obtained in the network after time shuffling the data. A link corresponds to a correlation between two signals $\geq r_c$ (shown $r_c = 0.8$. Results are similar for other values of $r_c$.) Actual results (red online) are greater than $4\sigma$ from the mean of the shuffled distribution.

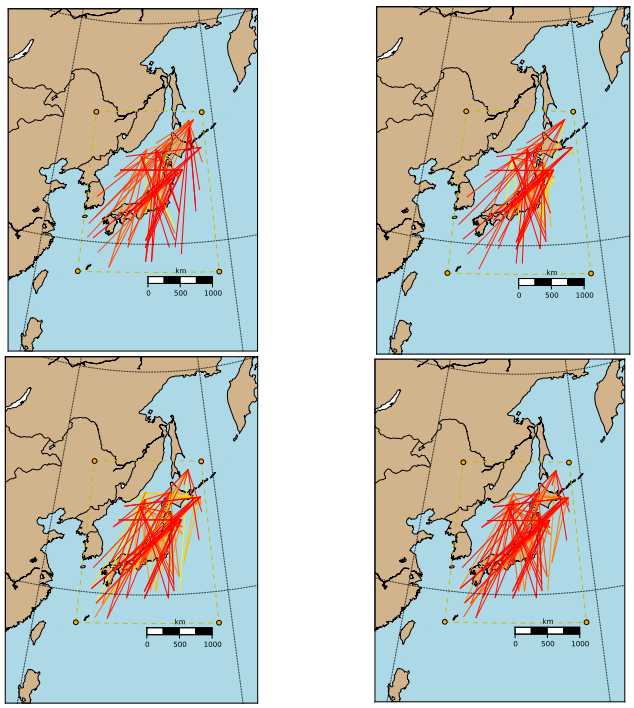

**Figure 5.4**: (Color online) Network links superimposed on a map of the Japanese archipelago, including Japan's main island Honshu. Note that links are anisotropic and primarily lie parallel to the principal axis of Honshu. Shown are links satisfying $r \geq r_c$ that are connected to high-degree nodes ($k \geq k_{min}$). Darker colors (red online) indicate stronger links (i.e. stronger correlations). Links shown satisfy (a) $r_c = 0.9$, $k_{min} = 5$, (b) $r_c = 0.8$, $k_{min} = 7$, (c) $r_c = 0.7$, $k_{min} = 8$, (d)$r_c = 0.5$, $k_{min} = 8$. These choices for $r_c$ and $k_{min}$ give approximately 70, 70, 90, and 90 links respectively.

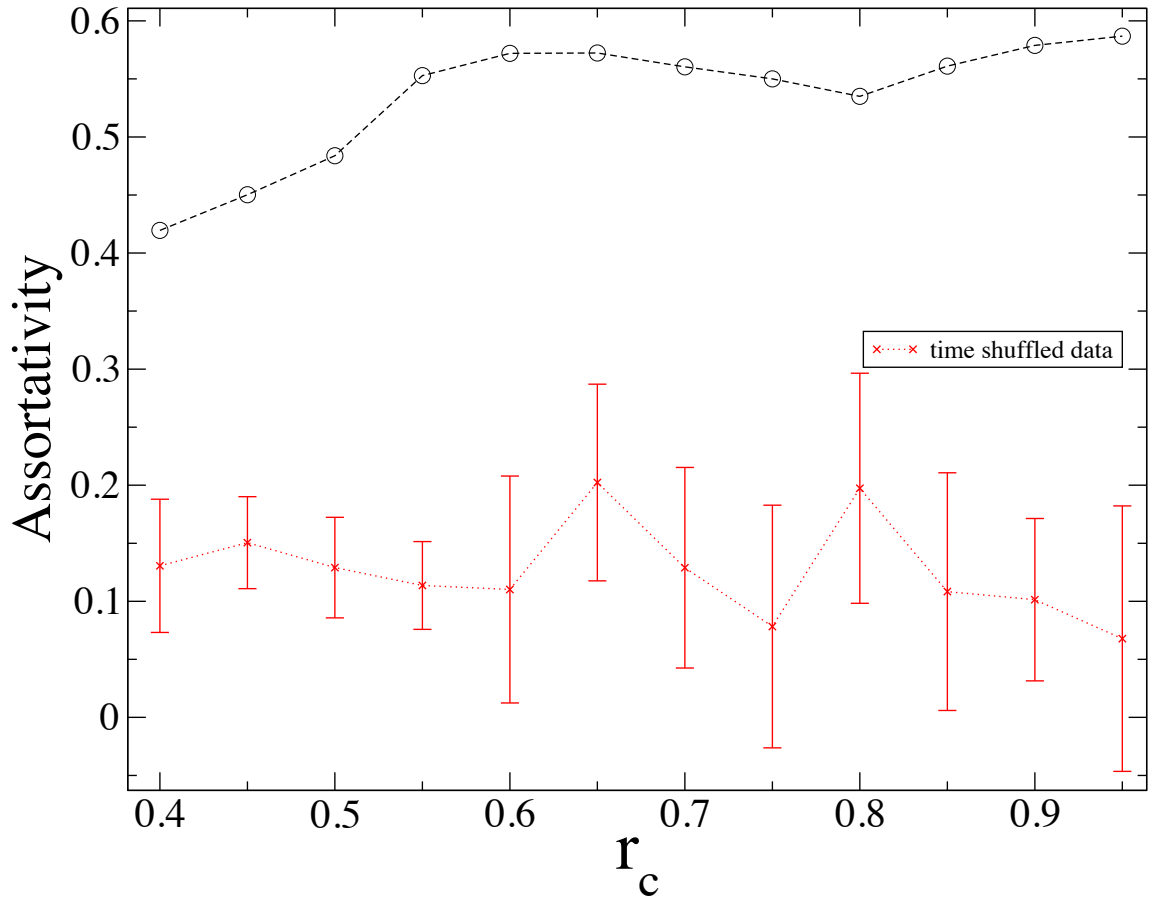


**Figure 5.5**: (Color online) Demonstration that earthquake networks are highly assortative for a wide range of $r_c$, generally increasing with $r_c$. Assortativity $> 0$ indicates that high-degree nodes tend to link to high-degree nodes and low-degree nodes tend to link to low-degree nodes. For comparison assortativity values obtained from networks using time-shuffled data demonstrate that these findings are not a finite-size effect or a result of spatial clustering (time-shuffling preserves location).

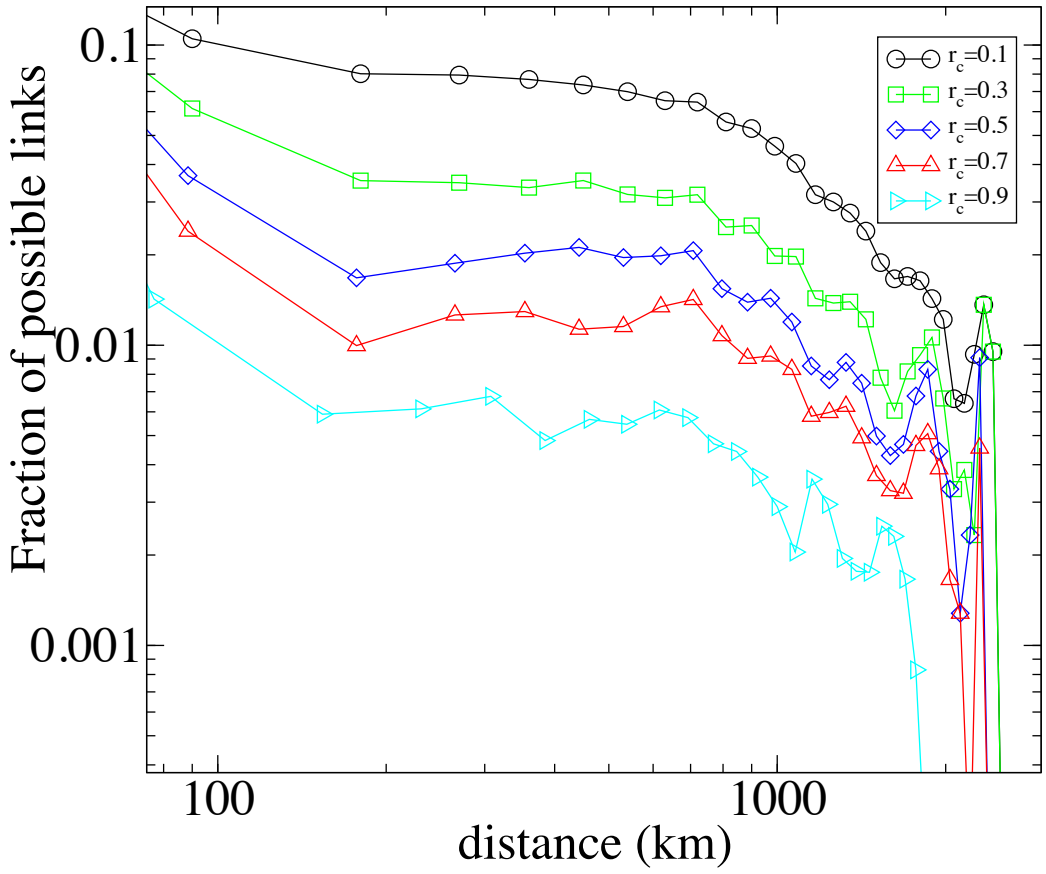


**Figure 5.6**: (Color online) Number of network links at a given distance as a fraction of how many links are geometrically possible at that distance, demonstrating that links have no characteristic length scale. Distances less than 100km have sparse statistics due to the coarseness of the spatial grid while distances greater than 2300km have sparse statistics due to the finite spatial extent of the catalog.

# Part III

# Asymmetry in power-law magnitude correlations

# Chapter 6

# Background

A familiar concept to many is the idea of a spatial fractal, as demonstrated, for example, by the famous Koch snowflake or Mandelbrot set. As noted in Chapter 2, seismic phenomena have also been shown to exhibit fractal behavior, both in the spatial occurrence of earthquakes and in the fault systems embedded within the earth's crust.

The defining characteristic of such a fractal is that the object looks the same at all scales, i.e. "zooming in" on any piece of the shape results in the reappearance of the same shape. This kind of self-similarity is a staple of complex systems. Complex systems have similar dynamics at a wide range of scales, which is necessarily fractal, nearly by definition.

This type of scale-invariance extends beyond phenomena that are spatial in nature to the temporal. While a simple signal is often characterized by two scales - the size characterizing the phenomenon itself and the characteristic size of the background noise - a complex signal like those of many financial indices and physiological data has no such abrupt cutoff at any scale.The observed fluctuations display the same behavior, no matter the scale at which the signal is examined. This self-similarity manifests in the form of power-law autocorrelations embedded in the signal and can be tested for using statistical techniques.

The outputs of a broad class of systems ranging from physical and biological, to social systems exhibit either long-range temporal or spatial correlations that can

be approximated by power laws [53, 57, 23]. A variety of studies have also found that different complex systems spanning finance [24], physiology [25], and seismology [26, 27] generate time series of increments, the absolute values (magnitudes) of which are power-law correlated. The correlation of these magnitudes, results in "clustering", where large increments are more likely to follow large increments and small increments are more likely to follow small increments. As a consequence of this, a simple random walk model fails to fully describe the data.

Random walks are commonly utilized in physics as a null model hypothesis to describe observed data. The data are compared to the statistical results of what can be thought of as the motion of a severely intoxicated person[28]. At each discrete time step the random walker has a probability $p$ to step to the right and therefore a probability $1-p$ to instead step to the left. In general, the sizes of the steps can also be allowed to vary. Using the drunk stumbler as a heuristic, we can then characterize the walk in terms of the total distance the walker is from the starting point, and calculate quantities like, on average, where we would expect the walker to be, how far the walker generally wanders over a certain time period, and the distance from the starting point can be expected to depend on time. As a typical feature of the random walk model, there is no memory in the walking process: the walker is so inebriated that e.g. taking a step left has no effect on his subsequent step so that each step is independent of the previous.

Because a variety of complex systems do display a memory, in that the result of one step causes changes in what the next step is likely to be, a simple random walk model misses important aspects of the data. To solve this problem, the random walk model is generalized to capture this clustering regularity. The random walker now continues to stumble left or right, but does so with a characteristic step size that depends on time.

Thus, long-range magnitude correlations in increments $x_i$ are usually modeled using a time-dependent standard deviation, $\sigma_i$ [24], commonly called volatility, describing this characteristic step size. $\sigma_i$ is defined as a linear combination of $N$ previous values of $|x_{i-n}|$, i.e. $\sigma_i = \Sigma_{n=1}^{N} a(n) x_{i-n}^2$, where $i$ refers to the $i$th term and $a(n)$ are statistical weights. $a(n)$ should be a decreasing function, e.g. power-law or exponential, since the most recent events (smaller $n$) intuitively contribute more than events

from the distant past. Such a model is referred to as an Autoregressive conditional heteroskedasticity (or ARCH) model[24].

Magnitude correlations were proposed in order to understand financial time series [24]. Magnitude correlations of many financial time series [29] are asymmetric with respect to increment sign, in that negative increments were more likely to be followed by increments of large magnitude and positive increments were more likely to be followed by increments with small magnitudes (i.e. "bad news" causes more volatility than "good news"). Such an observation should not be surprising given the findings in cognitive psychology that humans have a tendency to pay more attention to negative inputs and experiences than positive ones. Such an attentional bias can manifest in financial time series, where prices are influenced by human action[30][31].

If we are to model this asymmetry, the time-dependent standard deviation $\sigma_i$ we define must depend on both $x_{i-n}$ and $|x_{i-n}|$, to capture the dependence of both sign and magnitude. Since $\sigma_i$ must be positive, we can define $\sigma_i = \Sigma_n a(n)(||x_{i-n}|+\lambda x_{i-n}|)$, where $\lambda$ is a real parameter that acts as a measure of asymmetry. For $\lambda > 0$, positive increments $x_{i-1}$ are more likely to be followed by large increments $|x_i|$ (see Figs. 7.1 (a) and (b)), whereas for $\lambda < 0$, negative increments are more likely followed by large increments. $\lambda = 0$ reduces to the symmetric $\sigma_i$ above that has no dependence on the sign of the increment.

We ask if the concept of asymmetry in magnitude correlations is relevant to real-physical data. We first create a test allowing one to find if an observed asymmetry is statistically significant. We then propose a stochastic process in order to (i) further test significance, and (ii) model data as dependent on two parameters which characterize both the length of the power-law memory and its magnitude correlation asymmetry, parameters which we then demonstrate how to obtain. Finally, we apply our test to real-world physiological data to determine if there is statistically significant asymmetry in the magnitude correlations.

# Chapter 7

# Testing Statistical Significance

How would we know if an observed asymmetry is genuine and not due to a finite-size effect? For example, a *finite*-length time series generated by an independent and identically distributed (*i.i.d.*) (i.e., uncorrelated) process will exhibit a spurious asymmetry. To this end, we ask how large should the asymmetry be to become statistically significant? To answer this question, we generate *i.i.d.* series, and for each we calculate two sums, $S_+$ and $S_-$. The sum $S_+$ is the average of all the values $|x_i|$ preceded by positive $x_{i-1}$, while the sum $S_-$ is the average of all the values $|x_i|$ preceded by negative $x_{i-1}$. For an infinitely long *i.i.d.* time series, we expect $S_+ = S_-$, while finite length time series in general have $S_+ \neq S_-$.

We therefore define a test variable:

$$S \equiv S_+ - S_- \qquad . \tag{7.1}$$

What is the range $(-S_c, S_c)$ such that $S$ will fall in this range 95% of the time? To answer this question, we generate a large number of finite *i.i.d* time series, each with $N$ data points. For each time series we calculate $S$. We find on collecting all the $S$ values that $S$ follows a symmetrical probability distribution $P(S)$ centered at zero. By ranking the values $S$ from smallest to largest, we find a critical value $S_c$ for which there is probability 0.95 that the $S$ of a random uncorrelated series is between $(-S_c, S_c)$. By repeating the same procedure for a different number of data points, in Fig. 8.1 and in the inset, we find an almost perfect power-law fit relating the critical value $S_c$ to the number of data points with exponent $0.5 \pm 0.006$ in agreement with the Central Limit Theorem.

To find critical values for empirical series, we also use another approach of Ref. [27]. For a given series, we accomplish $10^4$ reshufflings, where each reshuffled time series is subtracted from the average and divided by its standard deviation. For each series, we calculate $S$ of Eq. 7.1. By ranking the values $S$ in ascending order, we find $S_c$ for which there is probability 0.95 that the $S$ is between $(-S_c, S_c)$. By using this approach, for subjects 2 and 8 we find $S_c = 0.019$ and $S_c = 0.021$, respectively.

We next argue that the interval $(-S_c, S_c)$ found for a given $N$ is a "litmus test" for significance. If the empirically calculated $S$ is found outside this interval, we consider the asymmetry statistically significant. We calculate the values of $S_c$ for various $N$ (Table 8.1 and Fig. 8.1).

Note that our test is model-independent — it measures asymmetry in magnitude correlations, but assumes neither the memory in correlations (long or short) nor the functional form of the correlation (e.g. power-law or exponential).

A concern is the possibility that in order to test significance of asymmetry in power-law magnitude correlations, we should find the intervals $(-S_c, S_c)$ not from *i.i.d.*, but from time series generated by symmetric magnitude correlations. To address this concern, we create a stochastic process characterized by asymmetric power-law correlations in the magnitudes $|x_i|$

$$x_i \;=\; \sigma_i \eta_i, \qquad \sigma_i = \sum_{n=1}^{\infty} a_n(\rho) \frac{||x_{i-n}| + \lambda x_{i-n}|}{\langle ||x_{i-n}| + \lambda x_{i-n}| \rangle}, \tag{7.2}$$

where $\rho \in (0, 0.5)$ and $\lambda \in (-1, 1)$ are free parameters, $\sigma_i$ is a time-dependent standard deviation, $a_n(\rho)$ are power-law distributed weights $a_n(\rho) = \Gamma(n-\rho)/(\Gamma(-\rho)\Gamma(1+n))$ chosen to generate power-law correlations in the magnitudes $|x_i|$. $\Gamma(x)$ denotes the Gamma function and $\eta_i$ denotes *i.i.d.* Gaussian random variables with mean $\langle \eta_i \rangle = 0$ and variance $\langle \eta_i^2 \rangle = 1$. The parameter $\rho$ controls the length of the power-law memory, whereas the parameter $\lambda$ controls the asymmetry in magnitude correlations. When $\lambda = 0$, the process of Eq. 7.2 reduces to a fractionally integrated autoregressive moving average (FIARCH) process with symmetric magnitude correlations [32] for which $\alpha = 0.5 + \rho$ [33], where $\alpha$ is the exponent found from detrended fluctuation analysis (DFA) [34]. We therefore call the process of Eq. 7.2 asymmetric FIARCH process (AFIARCH).

Because we include all previous increments in $\sigma_i$ of Eq. 7.2, our process is nec-

essarily *long-range* correlated. We can also create a *short-range* correlated process $x_i = \sigma_i \eta_i$ by including only the most recent increment, so $\sigma_i = (||x_{i-1}| + \lambda x_{i-1}|)$. In this paper, instead of $\ell = \infty$, in Eq. 7.2 we use the cutoff length $\ell = 500$.

By using the process of Eq. 7.2, we generate a number of time series and find that the magnitude correlations quantified by the DFA exponent practically do not depend on the parameter $\lambda$. To demonstrate this, Fig 7.1(b) shows DFA plots for two fixed values of $\rho$ and varying values of $\lambda$. We see that the DFA plots practically overlap, and that $\alpha = 0.5 + \rho$ holds, as for the symmetric FIARCH process Eq. (7.2 with $\lambda = 0$ [32]). Thus, the asymmetric term in Eq. 7.2 ($\lambda \neq 0$) practically does not affect the correlation pattern of the magnitude time series.

We next return to our goal of determining the statistical significance of asymmetry. We use the process of Eq. 7.2 with $\lambda = 0$ to generate a large number of time series for various values of $\rho$ and $N$. We then determine the test variable $S$ of Eq. 7.1 for each of these series. Ranking the values $S$ from smallest to largest we find a critical value $S_c$ for which there is probability of 0.95 that the $S$ from a finite symmetrically-defined series falls between $(-S_c, S_c)$. Varying both $\rho$ and $N$, in Fig. 8.1 we obtain four power-law fits relating the critical value $S_c$ and the number of data points $N$. As expected, the critical values for power-law correlated time series shown in Table 8.1 with $\rho = 0.1$ ("weak" power-law correlations) are practically the same as the critical values obtained for *i.i.d.* time series. However, the stronger the correlation, the larger the critical value $S_c$.

In order to estimate the parameter $\lambda$ characterizing the asymmetry of a time series, we employ the maximum likelihood estimation method [35]. One starts by deriving a likelihood function that is an expression for the probability of obtaining a given sample of $N$ known observations $(X_1, X_2, ..., X_N)$. We denote the probability of obtaining the *i-th* observation $X_i$ as $P(X_i)$. Then the probability $L$ of obtaining our particular $N$ observations is the product of the probability $P(X_i)$ to obtain each

$$L = \prod_{i=1}^{N} P(X_i). \tag{7.3}$$

To make further progress, we need to posit a form for $P(X_i)$. We assume the increments $X_i$ are normally distributed $P(X_i) = (2\pi\sigma_i^2)^{-1/2} \exp(X_i^2/2\sigma_i^2)$ with a mean 0 and characterized by a time-dependent variance $\sigma_i^2$ which depends on the past values

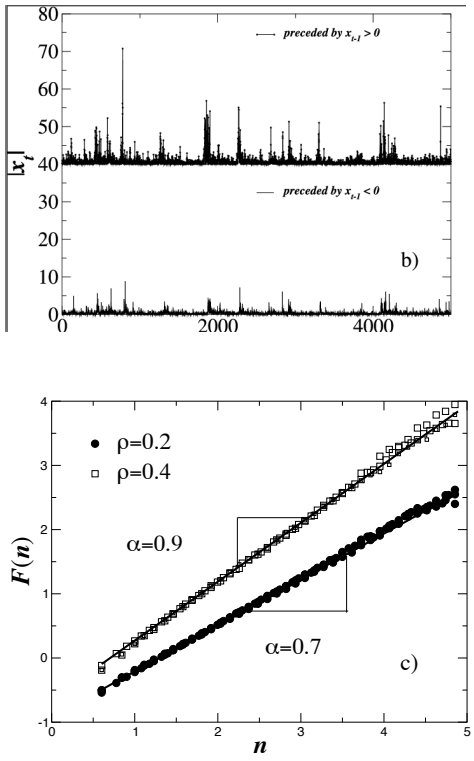


**Figure 7.1**: Asymmetry in magnitude correlations and detrended fluctuation function $F(n)$. (b) We show that the increments are larger for $x_{i-1} > 0$ (top curve, shifted upward for clarity) than for $x_{i-1} < 0$ (bottom curve) as a result of positive $\lambda$. The time series is obtained from numerical simulations of the process of Eq. 7.2 with $\lambda = 0.9$ and $\rho = 0.4$. (c) Detrended fluctuation function $F(n)$, where $n$ is a measure of window size, obtained from numerical simulations of process of Eq. 7.2 with $\lambda = 0.3, 0.6$, and 0.9 and $\rho = 0.2$ and 0.4. For asymptotically large values of $n$, each of the $F(n)$ curves can be approximated by a power law $F(n) \propto n^{\alpha}$ with scaling exponent $\alpha \approx 0.5 + \rho$ independent of the value of $\lambda$.

of $X_i$. In our case, all values of $\sigma_i$ and all values of $P(X_i)$ are characterized by only two adjustable parameters ($\rho$, $\lambda$). Substituting the previous $P(X_i)$ into Eq. 7.3, and taking the logarithm we obtain the log-likelihood function for the sample [35]

$$\ln L = -\frac{1}{2} N \ln(2\pi) - \sum_{i=1}^{N} \left[ \ln(\sigma_i) + \frac{1}{2} X_i^2 / \sigma_i^2 \right]. \tag{7.4}$$

where $\sigma_i$ is given by Eq. 7.2.

# Chapter 8

# Results

To illustrate the utility of the process of Eq. 7.2 for modeling real-world data, we next analyze a large electroencephalography (EEG) database [36] comprising records from 25 subjects randomly selected over a 6-month period at St. Vincent's University Hospital in Dublin [37]. EEG data are recorded every 0.8 s, so we obtain the number of data points $N$ between 22,000 and 30,000 (Table 8.2). Time series of EEG magnitudes exhibit power-law long-range correlated behavior [38, 39].

From each original time series we subtract the average. From Tables 8.1 and 8.2 we see that our test of Eq. 7.1 with probability 0.95 confirms the existence of asymmetry in magnitude correlations. The test for each subject is outside the range $(-S_c, S_c)$ for a given $N$. For example, for subject 02, characterized by $N = 28,000$ (close to 32,000 in Table 8.1) and $\rho = 0.27$ (close to 0.3 in Table 8.1) we find $S = -0.024$ that is outside the range we obtained for *i.i.d.* process $(-0.014, 0.014)$, process with symmetric power-law magnitude correlations $(-0.019, 0.019)$, and the approach of Ref. [27] $(-0.019, 0.019)$.

By minimizing Eq. 7.4, we estimate $\rho$ and $\lambda$ for each subject (Table 8.2) where we choose a normal pdf $P(x)$ for EEG data. Commonly one uses a normal pdf when using log-likelihood approach. In order to check if some other choice for $P(x)$ would be more appropriate, next we analyze pdf $P(x)$ of empirical data [36]. In Fig. 8.2 we see that for most of the empirical time series, $P(x)$ in the broad central region follows not normal, but Laplace distribution $P(X_i) = 1/\sqrt{2}\sigma \exp(-\sqrt{2}(x - \overline{x})/\sigma)$, where $\sigma$

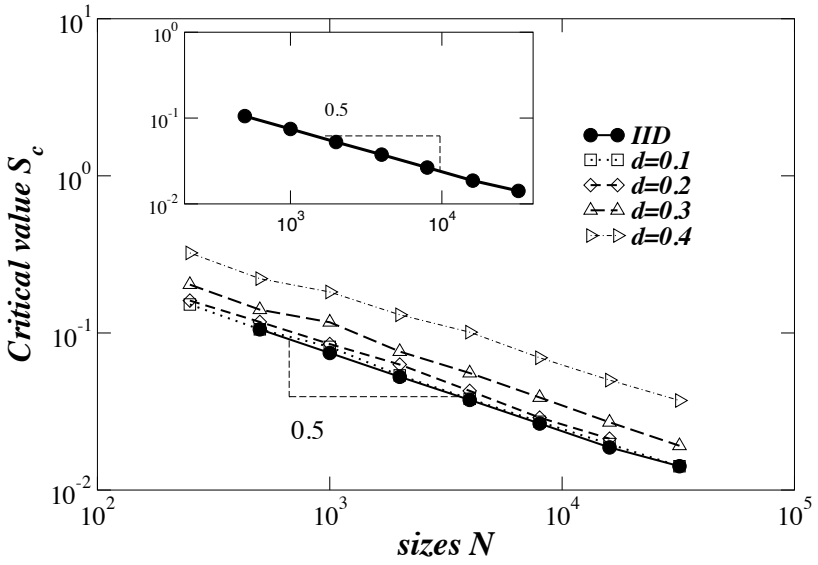


**Figure 8.1**: To utilize the test from Eq. (7.1), we show that the critical value $S_c$ follows a power law with respect to sample size $N$. In order to apply the test for empirical time series, we calculate values $S_c$ for different value of $N$. We generate $10^6$ *i.i.d.* time series for such $N$ and for each series we calculate the test $S$. We rank all the $S$ values from smallest to largest and find the $S_c$ for which there is probability 0.95 that the $S$ of a generated series is between $(-S_c, S_c)$. We repeat the same procedure for different values of $N$. We obtain a power law (inset) $S_c \sim AN^{-\alpha}$ between $S_c$ and $N$, where $\alpha = 0.5$ and $A = 2.16$. We repeat the procedure for $S_c$ values for the process of Eq. (7.2) when $\lambda = 0$ (FIARCH), a symmetric process in magnitude correlations. For four different values of $\rho$ we obtain power law relations between $S_c$ and $N$.

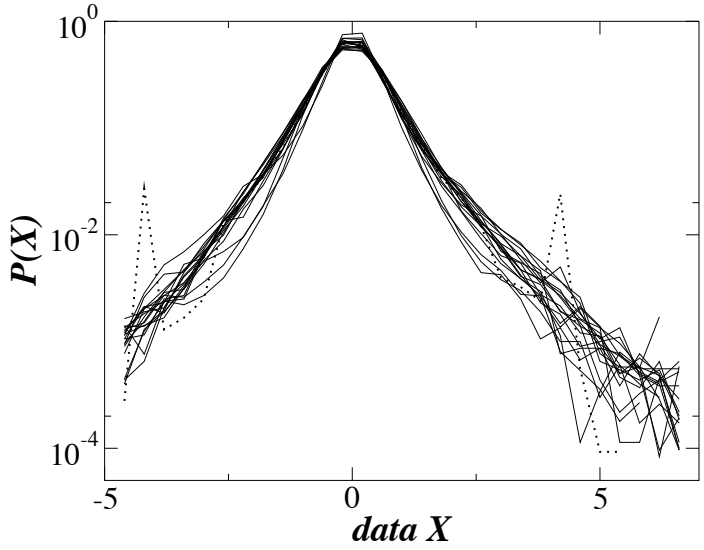


**Figure 8.2**: The pdf in the log-linear plot for each of 20 EEG time series out of total 25 time series [36] comprising records from 20 subjects. Each pdf approximately follows the Laplace distribution. We also show the pdf of subject 10 whose tails due to bumps deviate from the Laplace pdf.

is the standard deviation. For 5 subjects, $P(x)$ exhibit some bumps in the tail parts.

Next we follow the procedure of Eq. 7.3, but this time with Laplace $P(x)$. We find that the parameters $\rho$ and $\lambda$ change quantitatively, but not qualitatively (see Table 8.2) — the sign of $\lambda$ does not change by replacing normal $P(x)$ by Laplace $P(x)$.

To further test the statistical significance of asymmetry in magnitude correlations found in the data, we employ another process known as asymmetric GARCH (AGARCH) process [29]. This process is characterized by an exponentially decaying auto-correlation function

$$\sigma_i^2 = \omega + \alpha(|x_{i-1}| + \lambda_1 x_{i-1})^2 + \beta\sigma_{i-1}^2, \tag{8.1}$$

where $\alpha$, $\beta$, $\omega$, and $\lambda_1$ are the free AGARCH parameters and $\lambda_1$ is an asymmetry parameter similar to the one in Eq. 7.2. The last column in Table 8.2 shows our estimate for $\lambda_1$ (and two standard errors).

Note that the estimates $\lambda$ and $\lambda_1$ for AFIARCH and AGARCH, respectively, calculated for different subjects are very closely related. We obtain $\lambda = -0.029 + 1.232\lambda_1$, where 0.120 is the standard error of the slope coefficient. Differences are expected since the AFIARCH is characterized by power-law magnitude correlations (see Fig. (7.1(b)), while the AGARCH process is characterized by exponential magnitude correlations. From the results obtained for the AGARCH process, the asymmetry parameter $\lambda_1$ is statistically insignificant (within two standard errors) only for subjects 7, 8, 10, 12 and 21. Discrepancy between the results obtained from the test of Eq. 7.1 and the stochastic process of Eq. 8.1 is likely explained by the fact that the test of Eq. 7.1 measures only asymmetry in magnitude correlations, and does not assume either (i) the functional form of the magnitude correlations or (ii) their long-range nature, whereas our stochastic process imposes both.

From the analysis of individual time series, we conclude that magnitude correlations in observed physiological data exhibit significant asymmetry. However, universality is not confirmed. From the values of $\lambda$ in Table 8.2 obtained for different subjects, we calculate the average $\lambda$ — $\overline{\lambda} = -0.022 \pm 0.11$. The spread of values of the asymmetry parameter $\lambda$ suggests that the asymmetry does not show universality. However, based on the findings of [25], which showed statistical differences between healthy and diseased subjects, the average $\lambda$ and its standard deviation $\sigma$ might also show significant differences between diseased and healthy subjects. Consequently, the present analysis and proposed test may have potential to be useful for diagnostic purposes.

**Table 8.1**: Critical values $S_c$ obtained for an *i.i.d.* process and for the process of Eq. 7.2 with $\lambda = 0$. Due to finite-size effects, even these two processes may have non-zero values for $S$. In order to be considered significant asymmetry, we demand that the empirically calculated $S$ (see Table 8.2) is outside the interval $(-S_c, S_c)$.

| N | *i.i.d.* | $\rho = 0.1$ | $\rho = 0.2$ | $\rho = 0.3$ | $\rho = 0.4$ |
|---|---|---|---|---|---|
| 500 | 0.105 | 0.106 | 0.118 | 0.141 | 0.222 |
| 2,000 | 0.053 | 0.054 | 0.063 | 0.076 | 0.131 |
| 8,000 | 0.027 | 0.027 | 0.029 | 0.039 | 0.070 |
| 32,000 | 0.014 | 0.014 | 0.016 | 0.019 | 0.037 |

**Table 8.2**: Subjects shown in column 1 are designated as in Ref. [36]. We show only a few subjects. Statistics for all subjects available on request. Column 2 shows the number $N$ of data points, while column 3 displays the results obtained for the test of Eq. 7.1. In columns 4 and 5 are the estimates for $\rho$ and $\lambda$ of the AFIARCH process of Eq. 7.2 obtained after likelihood minimization of Eq. 7.4 with Gaussian pdf. In columns 6 and 7 are the AFIARCH estimates for $\rho$ and $\lambda$ with Laplace pdf. In the last column we show the AGARCH $\lambda_1$ estimate.

| subject | N | S | $\rho_G$ | $\lambda_G$ | $\rho_L$ | $\lambda_L$ | AGARCH $\lambda_1$ |
|---|---|---|---|---|---|---|---|
| 02 | 28,000 | -0.024 | 0.27 | 0.20 | 0.30 | 0.09 | $0.170 \pm 0.01$ |
| 06 | 30,000 | 0.015 | 0.24 | -0.02 | 0.22 | -0.01 | $-0.085 \pm 0.010$ |
| 08 | 24,000 | 0.070 | 0.28 | 0.02 | 0.27 | 0.03 | $-0.002 \pm 0.012$ |
| 10 | 24,000 | 0.028 | 0.24 | 0.10 | 0.26 | 0.13 | $0.023 \pm 0.022$ |
| 24 | 25,000 | 0.003 | 0.20 | -0.08 | 0.19 | -0.08 | $-0.062 \pm 0.012$ |
| 28 | 27,000 | 0.028 | 0.24 | -0.12 | 0.25 | -0.06 | $-0.115 \pm 0.010$ |

# Part IV

# Comparison between response dynamics in transition and developed economies

# Chapter 9

# Background: Volatility Asymmetry

The focus of econophysics is marrying economics to physics by importing techniques and concepts from the latter to the former. However, one must be cautious in doing so, since a number of generalizations that can be taken for granted in physics do not extend to economics.

i) A physics law discovered in the U.S. presumably holds universally over the entire Earth. By contrast, few expect such country invariance to hold in economics, since economics laws tend to depend on the wealth level of a country. For example, the hypothesis of the weak form of market efficiency [40], which assumes that stock prices at any future time cannot be predicted, holds in many large developed markets, even as evidence of violation has been found in ten transition (developing) economies in Eastern and Central Europe [43, 44]. Also, highly developed economies [46, 47, 48, 49] and those of different levels of aggregation (continents) [50] display power-law probability distributions in their price fluctuations. However, analysis of the Indian National Stock Exchange may instead show exponential distributions. [45]

ii) There is no guarantee that economics laws are time-independent, even for countries of a given level of wealth. Just how time-dependent economics laws are remains under investigation.

We seek here to explore (i) the extent to which economics laws depend on both level of economic development and (ii) time.

To review from Chapter 5, many complex systems exhibit temporal or spatial correlations that can be approximated by power-law scaling [53, 54, 55, 56, 57, 23, 58, 59],

and a range of stochastic models [60, 61, 62, 63] have been proposed to explain this scale invariance. Recent studies have reported that power-law correlations in empirical data are often characterized by a significant skewness or asymmetry in the distributions of increments. Examples include astrophysical data [64], genome sequences [65], respiratory dynamics [66], brain dynamics [67], heartbeat dynamics [68], turbulence [69], physical activities, finance [70, 71, 72], and geophysics weather data [73]. Besides power-law correlations in the increments, different complex systems exhibit power-law correlations in the absolute values of increments. Examples include finance [24, 74], physiology [25], and seismology [26, 27]. Applications for this phenomenon are particularly salient in finance because the absolute values measure the level of financial risk.

As was mentioned in Chapter 5, the autoregressive conditionally heteroscedastic (ARCH) process models such time series as a random walk with variable-sized steps. The size of these steps (called the "volatility") in the simple ARCH model is dependent only on the previous terms in the series[24]. The time dependence is thus captured by defining the volatility at a given point to be dependent on the previous increments in the series. The question arises of whether this volatility is dependent not only on the magnitude of preceding increments but also on their sign. Commonly, stockholders may not react equally to bad news (negative price increments) as compared to good news (positive price increments). Again, this observation corroborates what one would intuitively expect from findings in cognitive psychology, demonstrating the perceptual bias of humans to spend more attention on negative inputs than equally significant positive ones.[30][31]

Many extensions of original ARCH process [24] and its generalization (GARCH [74]) have been subsequently defined in order to incorporate such "asymmetry" [75, 76] (see Sec. III), which have shown significant asymmetry in a variety of developed markets [76, 24, 77, 78, 79, 80].

We ask how universal the phenomenon of volatility asymmetry is in global markets, particularly in transition economies, which often show different statistics than developed economies. For example, in contrast to the predominant behavior of financial time series of developed markets to exhibit only very short serial auto-correlations in price changes, financial time series of Central and Eastern European transition

economies exhibit longer memory [81, 82, 83, 44]. If the volatility asymmetry exists in transition economies, is it persistent or does it change over time? Likewise, what can we say about the persistence of volatility asymmetry in developed economies?

# Chapter 10

# Asymmetry model

Here, we extend the study [80] of volatility asymmetry to Central and Eastern European transition economies using a generalization of the ARCH process, finding that most of the indices under investigation also display statistically significant volatility asymmetry. Surprisingly, we find that such asymmetry is far more pronounced during the 2007-2009 world financial crisis than for the preceding eight years, indicating a greater universality of asymmetric market response during times of shared economic adversity.

Here, we investigate financial time series of index returns of eleven European transition economies of Central and Eastern Europe. Specifically, we analyze eleven stock market indices—PX (of the Czech Republic), BUX (Hungary), WIG (Poland), RTS (Russia), SKSM (Slovakia), SVSM (Slovenia), CRO (Croatia), NSEL30 (Lithuania), TALSE (Estonia), RIGSE (Latvia), and PFTS (Ukraine)—each corresponding to one of the eleven transition economies. As a representative of developed economies, we consider the U.S. stock market, analyzing three financial indices: the S&P500, NYSE, and NASDAQ. All data are recorded daily. As is common in economics, we define the relative price change (also called the return) in terms of the stock price $S(t)$ as

$$R_t \equiv \log S(t + \Delta t) - \log S(t), \tag{10.1}$$

where $\Delta t = 1$ corresponds to a time lag of one day. The increments used in time series are the returns with the average return subtracted so that the resulting series has a mean of zero.

To estimate the parameter $\gamma$ quantifying the volatility asymmetry of a time series, we employ the maximum likelihood estimation (MLE) method to ascertain which parameter values optimize the probability of a stochastic process to reproduce the observed time series. We start by deriving a likelihood function that is an expression for the probability of observing a given sample of $N$ known data points $(X_1, X_2, ..., X_N)$. We denote the probability of obtaining the *i-th* data point $X_i$ as $P(X_i)$. Then the probability $L$ of obtaining our particular $N$ data points is the product of the probability $P(X_i)$ to obtain each

$$L = \prod_{i=1}^{N} P(X_i). \tag{10.2}$$

The most widely used volatility processes are based on the ARCH approach. The GARCH(1,1) and ARCH($n$) processes, for example, have the volatility, or time-dependent standard deviation, expressed by the squares of the increments, a choice which necessarily loses information because it eliminates the ability to explore if the volatility has any dependence on the sign of an increment. In order to account for the possible asymmetric dependence on an increment's sign, different variants of GARCH processes have been proposed. Here we employ GJR GARCH(p,q) [75], a process that incorporates this asymmetry. In order to model long memory in volatility auto-correlations, the current volatility $\sigma_t$ depends on $p$ prior volatilities $\sigma_{t-i}$ and $q$ prior fluctuations $\epsilon_{t-i}$:

$$\epsilon_t \equiv R_t - \mu = \sigma_t \eta_t, \tag{10.3}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} (\alpha_i + \gamma T_{t-i}) \epsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2, \tag{10.4}$$

where $t$ stands for time, $\mu$ is the mean of the relative price change, $\sigma_t$ is the volatility, $\eta_t$ is random number chosen from a Gaussian distribution with a standard deviation of 1 and mean equal to 0. The coefficients $\alpha$ and $\beta$ are determined by MLE and $T_t = 1$ if $\epsilon_t < 0$, $T_t = 0$ if $\epsilon_t \geq 0$. The parameter $\gamma$ is expected to be positive, meaning that bad news (negative increments) enlarge volatility more than good news.

# Chapter 11

# Results

For the sake of simplicity, we set $p = q = 1$, as is commonly done, in all simulations for the process. Note that this defines the volatility as a short-term parameter because it results in Equation 10.4 only going back one term. We use the *gretl* package [84]. For each of the 11 different indices of transition economies, during approximately the 10.5-year period studied, we calculate the parameters of the GJR GARCH(1,1) process of Eqs. 10.3-10.4 and present our results in Table 11. In the GJR GARCH process the sum $\alpha + \beta$ gives a measure of the volatility persistence. The smaller this sum is, the longer the characteristic lifetime of the persistence. Based on these results, with the exception of the Lithuanian NSEL30 index, all the indices exhibit volatility persistence since the sum $\alpha + \beta$ is close to 1. In addition, we find the asymmetry parameter $\gamma$ to be statistically significant to within two standard deviations for all indices except the Slovakian SKSM and Estonian TALSE. We find the largest asymmetry parameter $\gamma$ for the Russian RTS and Lithuanian NSEL30 indices. These two markets are characterized by the largest kurtosis (i.e. heaviest tails, corresponding to a higher incidence of extreme events) and thus the largest volatility. In our study, this implies that greater asymmetry corresponds with larger volatility, as found for some other developed markets. For all the indices except SKSM and TALSE, we find that $\gamma$ is positive.

Next we ask whether the statistical properties concerning volatility asymmetry are homogeneous. For comparison, DNA chains are not homogeneous in correlations in

that long-range correlations exist only in intron-containing genes, and not in intron-less genes [34]. Hence, we ask if these statistical properties are more pronounced, for example, during market crashes and economic crisis.

To answer this question, we split the entire period, 12/31/98-07/10/09, into two subperiods: a "control" period, 12/31/98-01/01/07, and a "crash" period, 01/01/07-07/10/09, chosen to coincide with the world financial crisis. Note that the late-2000s recession began in the United States in December 2007 and it was announced in July 2009 that the recession may have ended. The recession has been followed by the global financial crisis. For each of 11 different indices, and for each subperiod, we estimate the GJR GARCH(1,1) process of Eqs. 10.3-10.4 and present the results for $\alpha+\beta$ and $\gamma$ in Table 11.2. First we note that the parameter $\alpha+\beta$ changes little during these two subperiods. Four indices—RTS, BUX, PX, SKSM, and NSEL30—for both subperiods exhibit significant volatility asymmetry. For five other indices—WIG, SVSM, RIGSE, and CRO—the control subperiod is characterized by no statistically significant volatility asymmetry, while the crash period is characterized by statistically significant volatility asymmetry. Note that the TALSE index exhibits no volatility asymmetry in either subperiod. We also find that for all indices, except the Russian index, the asymmetry parameter $\gamma$ estimated for the last $\approx$2.5 year period characterized by 2007-2009 global recession and severe market crash is larger than $\gamma$ estimated for the previous less volatile eight-year period. In summary, the last $\approx$2.5 years of the 2007-2009 world financial crisis are characterized by larger and statistically more significant volatility asymmetry than the previous eight years.

Additionally, we compare the persistence of the auto-correlations in the transition economies to that in developed economies by comparing the sum of the parameters $\alpha+\beta$. The smaller this sum is, the longer is the characteristic lifetime of the dependence. Table 11.2 shows that the parameter $\alpha+\beta$ responsible for persistence in auto-correlations, except for the CRO and NSEL30 indices, does not change much for different subperiods. Note that the $\beta$ parameter determines the weight applied to the previous volatility, whereas the $\alpha$ parameter determines the weight applied to the most recent news. In contrast to $\alpha+\beta$, the parameter $\gamma$, which controls the volatility asymmetry, changes substantially for different subperiods. In Fig. 11.1 we show how $\gamma$ estimated annually (by using $\approx$ 252 daily returns) for different indices changes over

time. Fig. 11.1(a) shows the annual variation of $\gamma$ for representative countries with statistically significant $\gamma$ for both subperiods. In Fig. 11.1(b) we show $\gamma$ vs year for countries with statistically significant $\gamma$ only for the crash subperiod. In Fig. 11.1(b) we find that $\gamma$ substantially changes from positive to negative values.

For reference, we also compare our results to those of well-developed markets, using the S&P 500, NASDAQ, and NYSE Composite indices. Applying our method to the S&P 500 for the last 20 years in one-year intervals, we find that $\gamma$ varies over time, and is always positive. As an interesting result we find that the *smallest* $\gamma$ values occur in 2002 and 2007-2009 corresponding to the *dot-com* bubble crash and current global recession respectively. We repeat our analysis, this time with two-year intervals on the S&P 500, and we also include the NASDAQ and NYSE Composite indices. Our results are shown in Fig. 11.2. We find the smallest $\gamma$ values for 1982-1983 and 1988-1989 periods, proximal to the 1982 recession and Black Monday in 1987, respectively. Restricting ourselves to the last decade, the smallest $\gamma$ values again occur in the time periods matching the *dot-com* crash and 2007-2009 recession. Due to sample variability, we expect the asymmetry parameter $\gamma$ to change over time. We find, however, that for the well-known US indices $\gamma$ tends to decrease during market crashes and economic crises, time periods characterized by their large volatility.

Our results contradict the suggestion given by Refs. [85, 86] that a decrease in overall volatility implies a decrease in asymmetry. However, our work is in agreement with Ref. [87] where opposite results were found analyzing Asia-Pacific Stock Index Returns. Ref. [87] found that high-volatility regimes (indicated by "fatter" tails returns) are associated with relatively low asymmetry. We therefore are in a position to confirm this finding for the leading US financial indices. The negative association between volatility and asymmetry is obvious during both the *dot-com* bubble crash and the 2007-2009 global recession.

**Table 11.1**: Estimates of GJR GARCH(1,1) with standard errors in parenthesis.

| index | GARCH $\alpha_1 + \beta_1$ | $\alpha_1$ | $\beta_1$ | $\gamma$ | Log-likelihood |
|---|---|---|---|---|---|
| RTS | 0.981 | 0.065 (0.005) | 0.905 (0.006) | 0.276 (0.034) | −5094 |
| BUX | 0.977 | 0.085 (0.008) | 0.885 (0.010) | 0.212 (0.030) | −4923 |
| WIG | 0.990 | 0.055 (0.005) | 0.932 (0.005) | 0.140 (0.035) | −4673 |
| SKSM | 0.996 | 0.037 (0.002) | 0.959 (0.002) | −0.021 (0.021) | −4318 |
| SVSM | 0.962 | 0.314 (0.016) | 0.648 (0.014) | 0.105 (0.018) | −2949 |
| PX | 0.979 | 0.116 (0.011) | 0.846 (0.013) | 0.234 (0.030) | −4520 |
| PFTS | 0.931 | 0.184 (0.009) | 0.748 (0.009) | 0.015 (0.012) | −5218 |
| NSEL30 | 0.821 | 0.184 (0.016) | 0.613 (0.021) | 0.260 (0.030) | −3365 |
| RIGSE | 0.956 | 0.219 (0.013) | 0.738 (0.012) | 0.075 (0.022) | −3817 |
| TALSE | 0.999 | 0.098 (0.005) | 0.910 (0.003) | −0.016 (0.013) | −3967 |
| CRO | 0.950 | 0.193 (0.014) | 0.752 (0.016) | 0.076 (0.022) | −2941 |

**Table 11.2**: For two subperiods 1998/12/31 - 2006/01/01 and 2007/01/01 - 2009/07/10 we estimate the GJR GARCH(1,1) process with standard errors in parenthesis.

| index | $\alpha_1 + \beta_1$ | $\gamma$ | $\alpha_1 + \beta_1$ | $\gamma$ |
|---|---|---|---|---|
| RTS | 0.946 | 0.332 (0.049) | 0.986 | 0.259 (0.053) |
| BUX | 0.962 | 0.177 (0.036) | 0.975 | 0.246 (0.068) |
| WIG | 0.992 | 0.026 (0.044) | 0.959 | 0.465 (0.201) |
| SKSM | 0.967 | −0.075 (0.040) | 0.999 | 0.139 (0.038) |
| SVSM | 0.893 | 0.015 (0.024) | 0.932 | 0.231 (0.045) |
| PX | 0.957 | 0.203 (0.042) | 0.975 | 0.257 (0.055) |
| PFTS | 0.865 | 0.007 (0.017) | 0.979 | 0.025 (0.028) |
| NSEL30 | 0.751 | 0.156 (0.053) | 0.692 | 0.308 (0.061) |
| RIGSE | 0.953 | −0.020 (0.026) | 0.927 | 0.275 (0.053) |
| TALSE | 0.999 | −0.035 (0.019) | 0.999 | 0.042 (0.024) |
| CRO | 0.788 | −0.087 (0.040) | 0.979 | 0.164 (0.041) |

**Table 11.3**: For two subperiods 1998/12/31 - 2006/01/01 (subscript 1) and 2007/01/01 - 2009/07/10 (subscript 2) we show the standard deviation and the GJR GARCH(1,1) estimation.

| index | $\sigma_1$ | $\sigma_2$ | $\gamma_1$ | $\gamma_2$ |
|---|---|---|---|---|
| RTSI | 1.469 | 3.002 | 0.3651 | 0.2547 |
| BUX | 1.467 | 2.155 | 0.1977 | 0.2278 |
| WIG | 1.357 | 1.723 | 0.0173 | 0.4994 |
| SKSM | 1.312 | 1.076 | -0.0832 | 0.0972 |
| SVSM | 0.657 | 1.573 | 0.0026 | 0.4255 |
| PX | 1.242 | 2.236 | 0.2232 | 0.2782 |
| PFTS | 1.660 | 2.164 | -0.050 | 0.0631 |
| NSEL30 | 0.835 | 1.779 | 0.2337 | 0.3461 |
| RIGSE | 1.555 | 1.570 | -0.0531 | 0.3251 |
| TALSE | 1.078 | 1.431 | -0.0558 | 0.0891 |
| CRO | 1.130 | 2.060 | -0.1078 | 0.2053 |
| DOWJ | 1.069 | 1.794 | 1.0170 | 0.4410 |
| SP500 | 1.110 | 1.968 | 0.9802 | 0.4475 |
| FTSE100 | 1.125 | 1.794 | 1.3776 | 0.4966 |

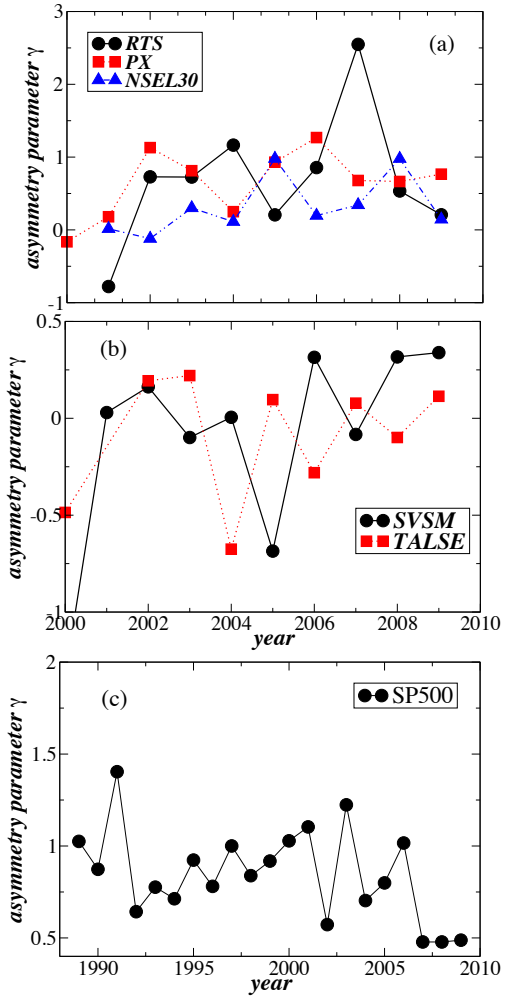


**Figure 11.1**: Changes of the volatility asymmetry parameter $\gamma$ each year over the 20-year period 1989-2009. (a) For transition economies $\gamma$ for both subperiods (crisis and control) changes over time. (b) The same, but for countries with statistically significant $\gamma$ only for the crisis subperiod. The parameter $\gamma$ substantially changes from positive to negative values. (c) As a representative for developed markets, we use the S&P 500 index. Over the last 20 years, $\gamma$ values vary over time, but $\gamma$ is always positive. The local minima for $\gamma$ values we obtain during *dot-com* bubble crash and during the 2007–2009 global crisis.

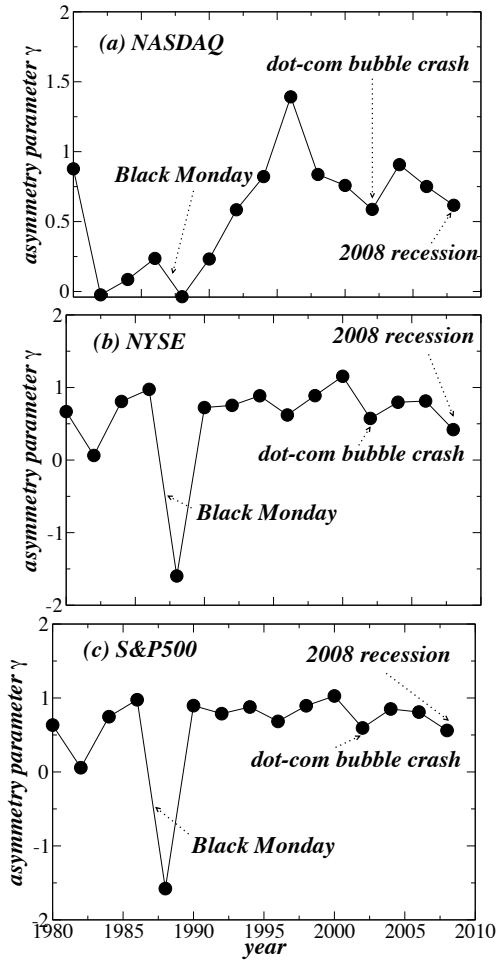


**Figure 11.2**: Changes of the volatility asymmetry parameter $\gamma$, calculated every two years over the 27-year period 1980-2008 for three developed markets: (a) NASDAQ, (b) NYSE, and (c) S&P500. Note the local minima for $\gamma$ values during Black Monday, the *dot-com* bubble crash, and during the 2007–2009 global crisis. The late-2000s recession began in the United States in December 2007.

# Part V

# Statistical laws governing fluctuations in word use from word birth to word death

# Chapter 12

# Background and Introduction

A number of arenas of competition demonstrate complexity in the form of scaling power laws. Again there is no characteristic size for many of the observed statistics because of the underlying hierarchy. This hierarchy has been shown e.g. in professional sports, academic careers, popular musical success, sexual activity, and in simple business competition, and follows from the so-called "Matthew effect", wherein the winners of a previous round of competition gain a probabilistic advantage for the next, creating a feedback loop in which a few players dominate (e.g. Babe Ruth, Google, Lady Gaga) and the majority play out at a more modest level, often obeying a power-law distribution among the big players[88, 89, 90, 91].

Within the context of language as a natural competitive arena between various words competing for reader attention, we extend the same line of investigation in competitive dynamics. We judge each word's success by how often the word is used relative to other words, an attribute which can convey information about the word's linguistic utility. For this approach to be meaningful, clearly, large amounts of data are necessary.

Several statistical laws describing the properties of word use, such as Zipf's law [92, 93, 94, 95, 96, 97] and Heaps' law [98, 99], have been exhaustively tested and modeled. However, since these laws are based on static snapshots aggregated over relatively small time periods and derived from corpora of relativity small size - from individual texts [92, 93] to collections of topical texts [94] and a relatively small snapshot of the British corpus [95] - little is known about the dynamical aspects of

language, including whether statistical regularities also occur in the time domain.

Do words, in all their breadth and diversity, display common patterns that are consistent with fundamental classes of competition dynamics? The data resulting from massive book digitization efforts allows us for the first time to probe this question in depth. Specifically, *Google Inc.* has recently unveiled a database of words, in seven languages, after having scanned approximately 4% of the world's books [100]. The massive project [101] allows for a novel view into the growth dynamics of word use and the birth and death processes of words in accordance with evolutionary selection laws [102]. Our focus is quantity $u_i(t)$, the number of uses of word $i$ in year $t$, which we regard as a proxy for the word's underlying linguistic value. Using the comprehensive *Google* dataset, we are able to analyze the growth of $u_i(t)$ in a systematic way for every word digitized over the 209-year time period 1800 – 2008 for the English, Spanish, and Hebrew text corpuses, which together comprise over $1 \times 10^7$ distinct words. This period spans the incredibly rich cultural history that includes several international wars, revolutions, and a number of paradigm shifts in technology. Here we use concepts from economics to gain quantitative insights into the role of exogenous (external) factors on the evolution of language, and we use methods from statistical physics to quantify the role of correlations both across words[106, 107, 108] and within a word itself. [103, 104, 105]

Since the number of books and the number of distinct words have grown dramatically over time (Fig. 12.1), we work mostly in terms of the *relative* word use, $f_i(t)$, (which we also refer to as the *"fitness"*) defined as the fraction of uses of word $i$ out of all word uses in the same year,

$$f_i(t) \equiv u_i(t)/N_u(t) \ , \tag{12.1}$$

where $N_u(t) \equiv \sum_{i=1}^{N_w(t)} u_i(t)$ is the total number of indistinct word uses digitized from books printed in year $t$, and $N_w(t)$ is the total number of distinct words digitized from books printed in year $t$. The relative use of a word depends on the intrinsic grammatical utility of the word (related to the number of "proper" sentences that can be constructed using the word), the semantic utility of the word (related to the number of meanings a given word can convey), and the context of the word's use. To quantify the dynamic properties of word prevalence at the micro- scale and its relation

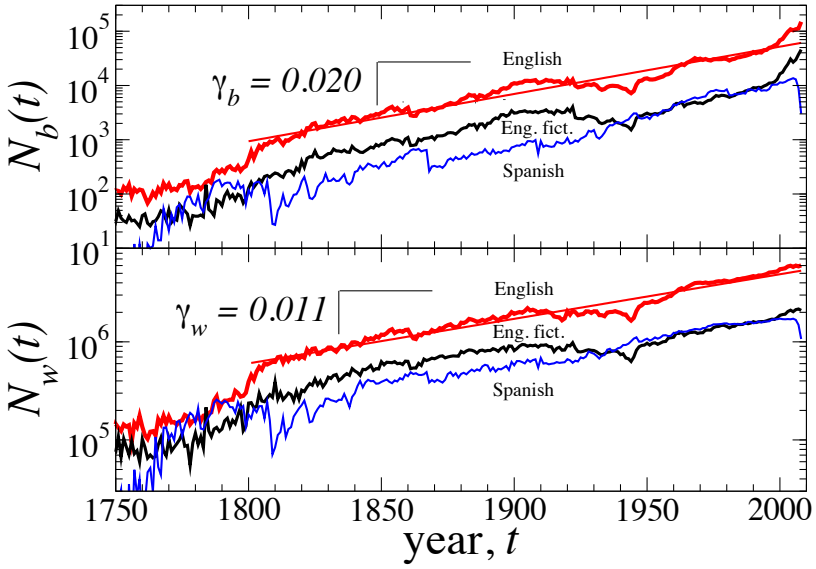


**Figure 12.1**: Since 1800, the number of books and the number of words has undergone approximately constant exponential growth with about 2% and 1% respectively.

to socio-political factors at the macro- scale, we analyze the logarithmic growth rate

$$r_i(t) \equiv \ln f_i(t+\Delta t) - \ln f_i(t) = \ln\left(\frac{f_i(t+\Delta t)}{f_i(t)}\right) , \tag{12.2}$$

a measure inspired by economic growth theory.

We treat words with equivalent meanings but with different spellings (e.g. color versus colour) as distinct words, since we view the competition among synonyms and alternative spellings in the linguistic arena as a key ingredient in complex evolutionary dynamics [109, 102]. A prime example of fitness-mediated evolutionary competition is the case of irregular and regular verb use in English. By analyzing the regularization rate of irregular verbs through the history of the English language, Lieberman et al. [110] show that the irregular verbs that are used more frequently are less likely to be overcome by their regular verb counterparts. Specifically, they find that the irregular verb death rate scales as the inverse square root of the word's relative use. Additionally, in neutral null models for the evolution of language [111], the fitness is the sole determining factor behind the survival capacity of the word in relation to its competitors.

We note also that the forces impacting the fitness have changed significantly over the years. With the advent spell-checkers in the digital era, words with spellings that

a spell-checker deems as standardized now receive a significant boost in their fitness at the expense of their "misspelled" counterparts. But not only "defective" words can die: even significantly used words can go extinct. For example, Fig. 13.5 shows three once-significant words, "Radiogram", "Roentgenogram" and "Xray", which competed in the linguistic arena for the majority share of nouns referring to what is now commonly known as an "Xray." The word"Roentgenogram" has since become extinct, even though it was the most common term for several decades in the 20th century. It is likely that two factors – (i) communication and information efficiency bias toward the use of shorter words [112] and (ii) the adoption of English as the leading global language for science – secured the eventual success of the word "Xray" by the year 1980.

# Chapter 13

# Results

**Quantifying the birth rate and the death rate of words.** Just as a new species can be born into an environment, a word can emerge in a language. Evolutionary selection laws can apply pressure on the sustainability of new words since there are limited resources (here books) for the use of words. Along the same lines, old words can be driven to extinction when cultural and technological factors limit the use of a word, in analogy to the environmental factors that can limit the survival capacity of a species by altering the ability of the species to obtain food in order to survive and reproduce.

We define the birth year $y_{0,i}$ as the year $t$ corresponding to the first instance of $f_i(t) \geq 0.05 f_i^m$, where $f_i^m$ is median word use $f_i^m = Median\{f_i(t)\}$ of a given word over its recorded lifetime in the *Google* database. Similarly, we define the death year $y_{f,i}$ as the last year $t$ during which the word use satisfies $f_i(t) \geq 0.05 f_i^m$. We use the relative word use threshold $0.05 f_i^m$ in order to avoid anomalies arising from extreme fluctuations in $f_i(t)$ over the lifetime of the word.

The significance of word births $\Delta_b(t)$ and word deaths $\Delta_d(t)$ for each year $t$ is related to the size of a language. We define the birth rate $r_b$ and death rate $r_d$ by normalizing the number of births and deaths in a given year $t$ to the total number of distinct words $N_w(t)$ recorded in the same year $t$, so that

$$r_b \equiv \Delta_b(t)/N_w(t) \ , \qquad (13.1)$$
$$r_d \equiv \Delta_d(t)/N_w(t) \ .$$

This definition yields a proxy for the rate of emergence and disappearance of words with respect to their individual lifetime use. We restrict our analysis to words with lifetime $T_i \geq 2$ years and words with a year of first recorded use $t_{0,i}$ that satisfies the criteria $t_{0,i} \geq 1700$, which biases for relatively new words in the history of a language.

Fig. 13.6 is a log-linear plot of the relative birth and death rates for the 208-year period 1800–2007. The modern era of publishing, which is characterized by more strict editing procedures at publishing houses and very recently computerized word editing with spell-checking technology, shows a drastic increase in the death rate of words, along with a recent decrease in the birth rate of new words. This phenomenon reflects the decreasing marginal need for new words, consistent with the sub-linear Heaps' law exponent calculated for all Google 1-gram corpora in [113].

Fig. 13.6 illustrates the current era of heightened word competition, demonstrated through an anomalous increase in the *death rate* of existing words and an anomalous decrease in the *birth rate* of new words. In the past 10–20 years, the total number of distinct words has significantly decreased, which we find is due largely to the extinction of both misspelled words and nonsensical print errors, and simultaneously, the decreased birth rate of new misspelled variations. This observation is consistent with both the decreasing marginal need for new words and also the broad adoption of automatic spell-checkers and corresponds to an increased efficiency in modern written language. Figs. 13.1 and 13.7 show that the birth rate is largely comprised of words with relatively large median $f_c$ (i.e. words that later became very popular) while the death rate is almost entirely comprised of words with relatively small median $f_c$ (words that never were very popular). Sources of error in the reported birth and death rates could be explained by OCR (optical character recognition) errors in the digitization process, which could be responsible for a certain fraction of the misspelled words. Also, the digitization of many books in the computer era does not require OCR transfer, since the manuscripts are themselves digital, and so there may be a bias resulting from this recent paradigm shift. Nevertheless, many of the trends we observe are consistent with the trajectories that extend back several hundred years.

Complementary to the death of old words is the birth of new words, which are commonly associated with new social and technological trends. Such topical words in modern media can display long-term persistence patterns analogous to earthquake

shocks [114, 115], and can result in a new word having larger fitness than related "out-of-date" words (e.g. log vs. blog, memo vs. email). Here we show that a comparison of the growth dynamics between different languages can also illustrate the local cultural factors (e.g. national crises) that influence different regions of the world. Fig. 13.8 shows how international crisis can lead to globalization of language through common media attention. Notably, such global factors can perturb the participating languages (here considered as arenas for word competition), while minimally affecting the nonparticipating regions, e.g. the Spanish speaking countries during WWII, see Fig. 13.8(a). Furthermore, we note that the English corpus and the Spanish corpus are the collections of literature from several nations, whereas the Hebrew corpus is more localized.

**The lifetime trajectory of words.** Between birth and death, one contends with the interesting question of how the use of words evolve when they are "alive". We focus our efforts toward quantifying the relative change in word use over time, both over the word lifetime and throughout the course of history. In order to analyze separately these two time frames, we select two sets of words: (i) relatively new words with "birth year" $t_{0,i}$ later than 1800, so that the relative age $\tau \equiv t - t_{0,i}$ of word $i$ is the number of years after the word's first occurrence in the database, and (ii) relatively common words, typically with $t_{0,i}$ prior to 1800. We analyze dataset (i) words, summarized in Table 13.1, so that we can control for properties of the growth dynamics that are related to the various stages of a word's life trajectory (e.g. an "infant" phase, an "adolescent" phase, and a "mature" phase). For comparison, we also analyze dataset (ii) words, summarized in Table 13.2, which are typically in a stable mature phase. We select the relatively common words using the criterion $\langle f_i \rangle \geq f_c$, where $\langle f_i \rangle$ is the average relative use of the word $i$ over the word's lifetime $T_i$, and $f_c$ is a cutoff threshold which we list in Table 13.2. In Table 13 we summarize the entire data for the 209-year period 1800–2008 for each of the four *Google* language sets analyzed.

Modern words typically are born in relation to technological or cultural events, such as "Antibiotics." We ask if there exists a characteristic time for a word's general acceptance. In order to search for patterns in the growth rates as a function of relative word age, for each new word $i$ at its age $\tau$, we analyze the "use trajectory"

$f_i(\tau)$ and the "growth rate trajectory" $r_i(\tau)$. So that we may combine the individual trajectories of words of varying prevalence, we normalize each $f_i(\tau)$ by its average $\langle f_i \rangle = \sum_{\tau=1}^{T_i} f_i(\tau)/T_i$ over the word's entire lifetime, obtaining a normalized use trajectory $f_i'(\tau) \equiv f_i(\tau)/\langle f_i \rangle$. We perform the analogous normalization procedure for each $r_i(\tau)$, normalizing instead by the growth rate standard deviation $\sigma[r_i]$, so that $r_i'(\tau) \equiv r_i(\tau)/\sigma[r_i]$ (see SI).

Since some words will die and other words will increase in use as a result of the standardization of language, we hypothesize that the average growth rate trajectory will show large fluctuations around the time scale for the transition of a word into regular use. In order to quantify this transition time scale, we create a subset $\{i \,|T_c\}$ of word trajectories $i$ by combining words that meets an age criteria $T_i \geq T_c$. Thus, $T_c$ is a threshold to distinguish words that were born in different historical eras and which have varying longevity. For the values $T_c = 25, 50, 100$, and 200 years, we select all words that have a lifetime longer than $T_c$ and calculate the average and standard deviation for each set of growth rate trajectories as a function of word age $\tau$. In Fig. 13.9 we plot $\sigma[r_i'(\tau|T_c)]$ which shows a broad peak around $\tau_c \approx 30$–$50$ years for each $T_c$ subset. Since we weight the average according to $\langle f_i \rangle$, we conjecture that the time scale $\tau_c$ is associated with the characteristic time for a new word to reach sufficiently wide acceptance that the word is included in a typical dictionary. The results of computing the mean first passage time to the critical frequency $f_c$ (i.e. the average time a word requires to achieve a critical amount of usage from its birth year) corroborate this conjecture (Fig. 13.9).

**Empirical laws governing the growth rates of word use.** How much do the growth rates vary from word to word? The answer to this question can help distinguish between candidate models for the evolution of word utility. Hence, we analyze the probability density function (pdf) for the normalized growth rates $R \equiv r_i'(\tau)/\sigma[r'(\tau|T_c)]$ so that we can combine the growth rates of words of varying ages. The empirical pdf $P(R)$ shown in Fig. 13.10 is remarkably symmetric and is centered around $R \approx 0$, just as is found for the growth rates of institutions governed by economic forces [116, 117, 118, 119]. Since the $R$ values are normalized and detrended according to the age-dependent standard deviation $\sigma[r'(\tau|T_c)]$, the standard deviation by construction is $\sigma(R) = 1$.

A candidate model for the growth rates of word use is the Gibrat proportional growth process [118], which predicts a Gaussian distribution for $P(R)$. However, we observe the "tent-shaped" pdf $P(R)$ which is a double-exponential or Laplace distribution, defined as

$$P(R) \equiv \frac{1}{\sqrt{2}\sigma(R)} \exp[-\sqrt{2}|R - \langle R\rangle|/\sigma(R)] \ . \tag{13.2}$$

Here the average growth rate $\langle R\rangle$ has two properties: (a) $\langle R\rangle \approx 0$ and (b) $\langle R\rangle \ll \sigma(R)$. Property (a) arises from the fact that the growth rate of distinct words is quite small on the annual basis (the growth rate of books in the Google English database is $\gamma_w \approx 0.011$ calculated in [113]) and property (b) arises from the fact that $R$ is defined in units of standard deviation. The Laplace distribution predicts a pronounced excess number of very large events compared to the standard Gaussian distribution. For example, comparing the likelihood of events above the $3\sigma$ event threshold, the Laplace distribution displays a five-fold excess in the probability $P(|R - \langle R\rangle| > 3\sigma)$, where $P(|R - \langle R\rangle| > 3\sigma) = \exp[-3\sqrt{2}] \approx 0.014$ for the Laplace distribution, whereas $P(|R - \langle R\rangle| > 3\sigma) = \mathrm{Erfc}[3/\sqrt{2}] \approx 0.0027$ for the Gaussian distribution. The large $R$ values correspond to periods of rapid growth and decline in the utility of words during the crucial "infant" and "adolescent" lifetime phases. In Fig. 13.10(b) we also show that the growth rate distribution $P(r')$ for the relatively common words comprising dataset (ii) is also well-described by the Laplace distribution.

For hierarchical systems consisting of units each with complex internal structure [120] (e.g. a given country consists of industries, each of which consists of companies, each of which consists of internal subunits), a non-trivial scaling relation between the standard deviation of growth rates $\sigma(r|S)$ and the system size $S$ has the form

$$\sigma(r|S_i) \sim S_i^{-\beta} \ . \tag{13.3}$$

The theoretical prediction in [120, 121] that $\beta \in [0, 1/2]$ has been verified for several economic systems, with empirical $\beta$ values typically in the range $0.1 < \beta < 0.3$ [121].

Since different words have varying lifetime trajectories as well as varying relative utilities, we now quantify how the standard deviation $\sigma(r|S_i)$ of growth rates $r$ depends on the cumulative word frequency

$$S_i \equiv \sum_{\tau=1}^{T_i} f_i(\tau) \tag{13.4}$$

of each word. To calculate $\sigma(r|S_i)$, we group words by $S_i$ and then calculate the standard deviation $\sigma(r|S_i)$ of the growth rates of words for each group. Fig. 13.11(b) shows scaling behavior consistent with Eq. 13.3 for large $S_i$, with $\beta \approx 0.10 - 0.21$ depending on the corpus. A positive $\beta$ value means that words with larger cumulative word frequency have smaller annual growth rate fluctuations. The emergent scaling is surprising, given the fact that words do not have internal structure, yet still display the analogous growth patterns of larger economically-driven institutions that do have complex internal structure. To explain this within our framework of words as analogs of economic entities, we hypothesize that the analog to the subunits of word use are the books in which the word appears. Hence, $S_i$ is proportional to the number of books in which word $i$ appears. As a result, we find $\beta$ values that are consistent with nontrivial correlations in word use between books. This phenomenon may be related to the fact that books are topical [94], and that book topics are correlated with cultural trends.

**Quantifying the long-term cultural memory.** Recent theoretical work [122] shows that there is a fundamental relation between the size-variance exponent $\beta$ and the Hurst exponent $H$ which quantifies the auto-correlations in a stochastic time series. The unexpected relation $\langle H \rangle = 1 - \beta > 1/2$ (corresponding to $\beta < 1/2$) indicates that the temporal long-term persistence, whereby on average large values are followed immediately by large values and smaller values followed by smaller values, can manifest in non-trivial $\beta$ values (i.e. $\beta \neq 0$ and $\beta \neq 0.5$). Thus, the $f_i(\tau)$ of common words with large $S_i$ display strong positive correlations and have $\beta$ values that cannot be explained by a either a Gibrat proportional growth, which predicts $\beta = 0$, or a Yule-Simon Urn model, which predicts $\beta = 0.5$.

To test this connection between memory ($H \neq 1/2$) and size-variance scaling ($\beta < 1/2$), we calculate the Hurst exponent $H_i$ for each time series belonging to the more relatively common words analyzed in dataset (ii) using detrended fluctuation analysis (DFA) [123, 124, 122]. We plot the relative use time series $f_i(t)$ for the words "polyphony," "Americanism," "Repatriation," and "Antibiotics" in Fig. 13.2A, along with DFA curves (see SI section) from which $H$ is derived in Fig. 13.2B. The $H_i$ values for these four words are all significantly greater than $H_r = 0.5$, which is the

expected Hurst exponent for a stochastic time series with no temporal correlations. In Fig. 13.3 we plot the distribution of $H_i$ values for the English fiction corpus and the Spanish corpus. Our results are consistent with the theoretical prediction $\langle H \rangle = 1 - \beta$ established in [122] relating the variance of growth rates to the underlying temporal correlations in each $f_i(t)$. This relation shows that the complex evolutionary dynamics we observe for words use growth is fundamentally related to the dynamics of cultural topic formation [125, 105, 114, 115] and dynamic bursting [126, 127].

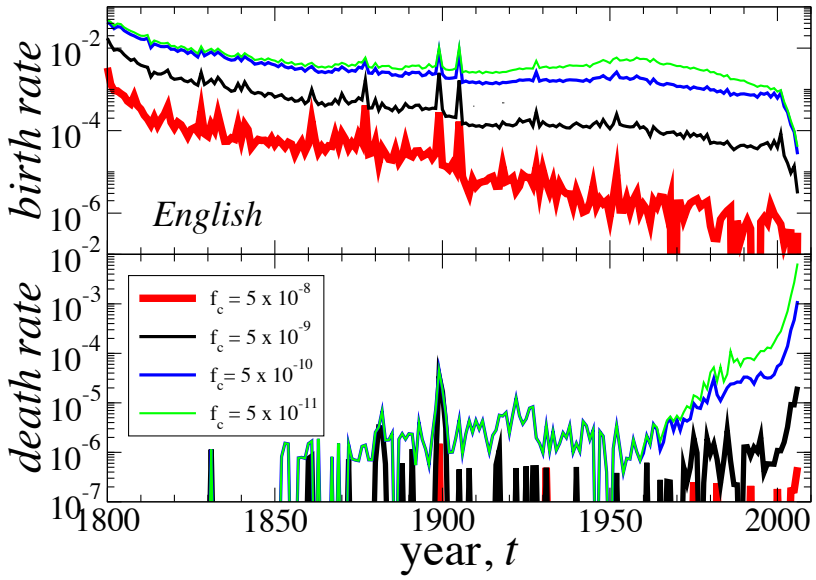


**Figure 13.1**: **The birth and death rates of a word depends on the relative use of the word.** For the English corpus, we calculate the birth and death rates for words with median lifetime relative use $\text{Med}(f_i)$ satisfying $\text{Med}(f_i) > f_c$. The difference in the birth rate curves corresponds to the contribution to the birth rate of words in between the two $f_c$ thresholds, and so the small difference in the curves for small $f_c$ indicates that the birth rate is largely comprised of words with relatively large $\text{Med}(f_i)$. Consistent with this finding, the largest contribution to the death rate is from words with relatively low $\text{Med}(f_i)$. By visually inspecting the lists of dying words, we confirm that words with large relative use rarely become completely extinct (see Fig. 13.5 for a counterexample word "Roentgenogram" which was once a frequently used word, but has since been eliminated due to competitive forces with other high-fitness competitors).

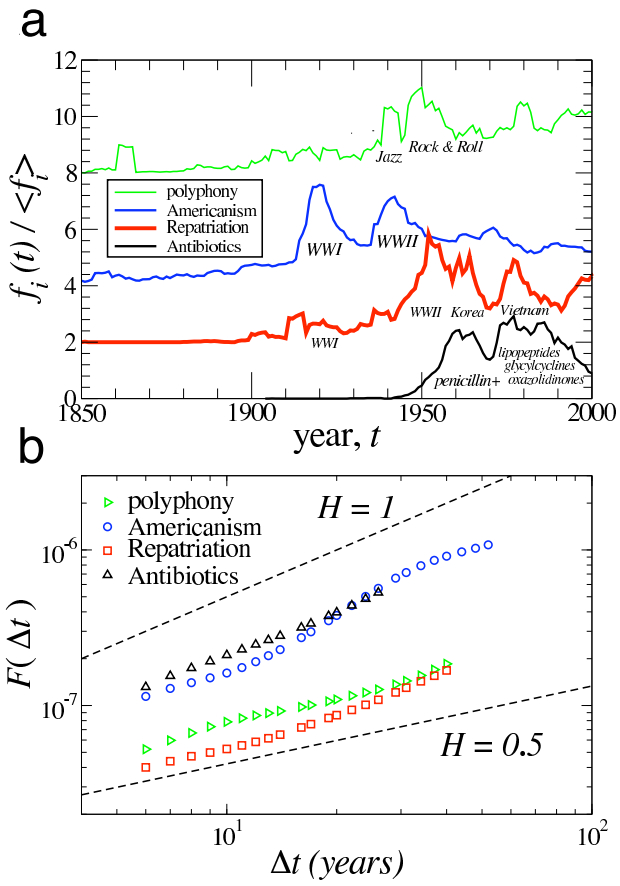


**Figure 13.2**: **Measuring the social memory effect using the trajectories of single words.** **(a)** Four example $f_i(t)$, given in units of the average use $\langle f_i \rangle$, show bursting of use as a result of social and political "shock" events. We choose these four examples based on their relatively large $H_i > 0.5$ values. The use of "polyphony" in the English corpus shows peaks during the eras of jazz and rock and roll. The use of "Americanism" shows bursting during times of war, and the use of "Repatriation" shows an approximate 10-year lag in the bursting after WWII and the Vietnam War. The use of the word "Antibiotics" is related to technological advancement. The top 3 curves are vertically displaced by a constant so that the curves can be distinguished. **(b)** We use detrended fluctuation analysis (DFA) to calculate the Hurst exponent $H_i$ for each word to quantify the long-term correlations ("memory") in each $f_i(t)$ time series. Fig. 13.3 shows the probability density function $P(H)$ of $H_i$ values calculated for the relatively common words found in English fiction and Spanish, summarized in Table 13.2.

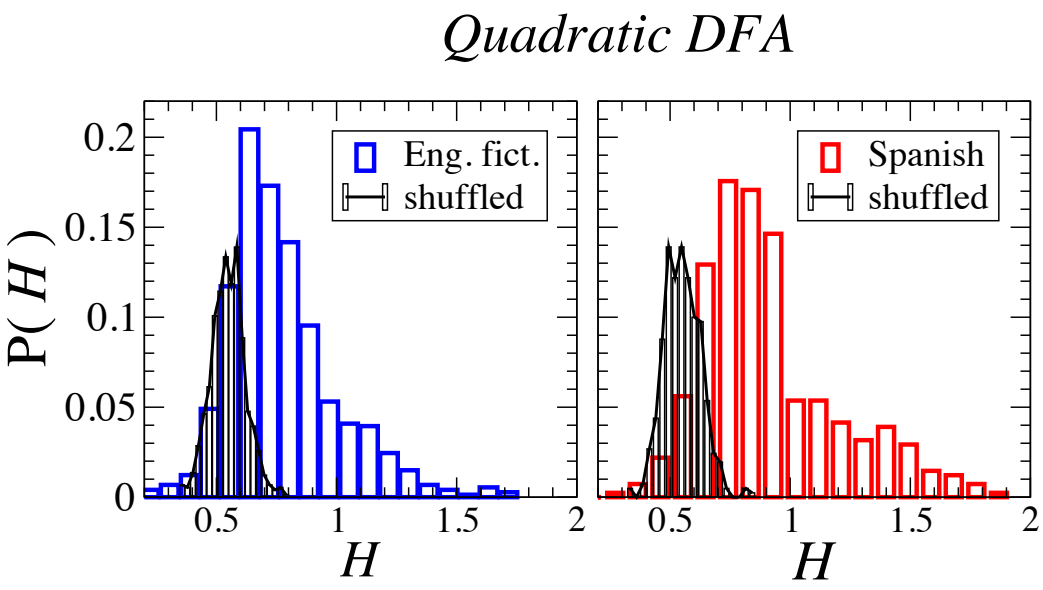


**Figure 13.3**: **Hurst exponent indicates strong correlated bursting in word use**. Results of detrended fluctuation analysis (DFA)[123, 124, 122] on the common [dataset (ii)] words analyzed in Fig. 13.10(b) show strong long-term memory with positive correlations ($H > 0.5$), indicating strong correlated bursting in the dynamics of word use, possibly corresponding to historical, social, or technological events. We calculate $\langle H_i \rangle \pm \sigma = 0.77 \pm 0.23$ (Eng. fiction) and $\langle H_i \rangle \pm \sigma = 0.90 \pm 0.29$ (Spanish). The size-variance $\beta$ values calculated from the data in Fig. 13.11 confirm the theoretical prediction $\langle H \rangle = 1 - \beta$. Fig. 13.11 shows that $\beta_{Eng.fict} \approx 0.21 \pm 0.01$ and $\beta_{Spa.} \approx 0.10 \pm 0.01$. For the shuffled time series, we calculate $\langle H_i \rangle \pm \sigma = 0.55 \pm 0.07$ (Eng. fiction) and $\langle H_i \rangle \pm \sigma = 0.55 \pm 0.08$ (Spanish), which are consistent with time series that lack temporal ordering (memory).

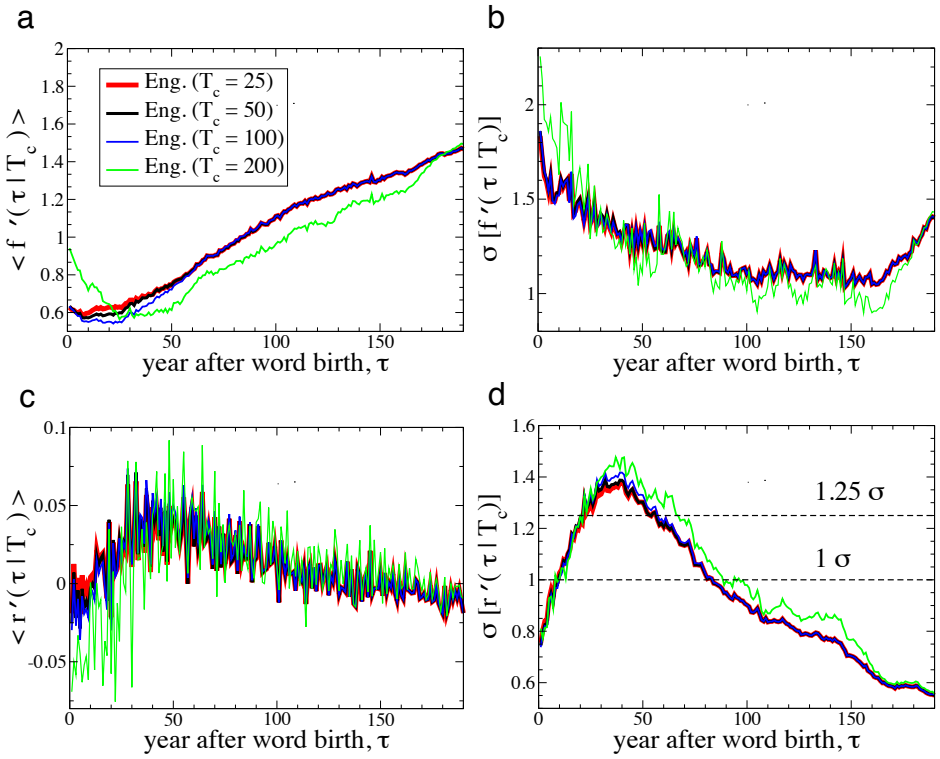


**Figure 13.4**: **Statistical laws for the growth trajectories of new words.** The "trajectory" of a words gives the word's popularity over its life. We show the word trajectories for dataset (i) words in the English corpus, although the same qualitative results hold for the other languages analyzed. $T_c$ denotes the lower bound on a word's lifetime (i.e. $T_i \geq T_c$), so that two trajectories calculated using different thresholds $T_c^{(1)}$ and $T_c^{(2)}$ only vary for $\tau < Max[T_c^{(1)}, T_c^{(2)}]$. The average is weighted according to $\langle f_i \rangle$. (a) The relative use increases with time, consistent with the definition of the weighted average which biases towards words with large $\langle f_i \rangle$. For words with large $T_i$, the trajectory has a minimum around $\tau \approx 40$ years, possibly reflecting the amount of time it takes to reach a critical fitness threshold of competition. (b) The variations in $\langle f(\tau|T_c) \rangle$ decrease with time reflecting the transition from the insecure "infant" phase to the more secure "adult" phase in the lifetime trajectory. (c) The average growth trajectory is qualitatively related to the logarithmic derivative of the curve in panel (a), and confirms that the region of largest positive growth is $\tau \approx 30$–$50$ years. (d) The variations in the average trajectory are largest for $30 \lesssim \tau \lesssim 50$ years and are larger than 1.0 $\sigma$ for $10 \lesssim \tau \lesssim 80$ years. Evidence shown in Fig. 13.9 supports that this is the time period for a word to be accepted into a standard dictionary.

**Table 13.1**: Summary of annual growth trajectory data for varying threshold $T_c$, and $s_c = 0.2$, $Y_0 \equiv 1800$ and $Y_f \equiv 2008$.

| Corpus, | Annual growth $R(t)$ data | | | | | |
|---|---|---|---|---|---|---|
| (1-grams) | $T_c(years)$ | $N_t(words)$ | % (of all words) | $N_R(values)$ | $\langle R \rangle$ | $\sigma[R]$ |
| English | 25 | 302,957 | 4.1 | 31,544,800 | $2.4 \times 10^{-3}$ | 1.00 |
| English fiction | 25 | 99,547 | 3.8 | 11,725,984 | $-3.0 \times 10^{-3}$ | 1.00 |
| Spanish | 25 | 48,473 | 2.2 | 4,442,073 | $1.8 \times 10^{-3}$ | 1.00 |
| Hebrew | 25 | 29,825 | 4.6 | 2,424,912 | $-3.6 \times 10^{-3}$ | 1.00 |
| English | 50 | 204,969 | 2.8 | 28,071,528 | $-1.7 \times 10^{-3}$ | 1.00 |
| English fiction | 50 | 72,888 | 2.8 | 10,802,289 | $-1.7 \times 10^{-3}$ | 1.00 |
| Spanish | 50 | 33,236 | 1.5 | 3,892,745 | $-9.3 \times 10^{-4}$ | 1.00 |
| Hebrew | 50 | 27,918 | 4.3 | 2,347,839 | $-5.2 \times 10^{-3}$ | 1.00 |
| English | 100 | 141,073 | 1.9 | 23,928,600 | $1.0 \times 10^{-4}$ | 1.00 |
| English fiction | 100 | 53,847 | 2.1 | 9,535,037 | $-8.5 \times 10^{-4}$ | 1.00 |
| Spanish | 100 | 18,665 | 0.84 | 2,888,763 | $-2.2 \times 10^{-3}$ | 1.00 |
| Hebrew | 100 | 4,333 | 0.67 | 657,345 | $-9.7 \times 10^{-3}$ | 1.00 |
| English | 200 | 46,562 | 0.63 | 9,536,204 | $-3.8 \times 10^{-3}$ | 1.00 |
| English fiction | 200 | 21,322 | 0.82 | 4,365,194 | $-3.5 \times 10^{-3}$ | 1.00 |
| Spanish | 200 | 2,131 | 0.10 | 435,325 | $-3.1 \times 10^{-3}$ | 1.00 |
| Hebrew | 200 | 364 | 0.06 | 74,493 | $-1.4 \times 10^{-2}$ | 1.00 |

**Table 13.2**: Summary of data for the relatively common words that meet the criterion that their average word use $\langle f_i \rangle$ over the entire word history is larger than a threshold $f_c$, defined for each corpus. In order to select relatively frequently used words, we use the following three criteria: the word lifetime $T_i \geq 10$ years, $1800 \leq t \leq 2008$, and $\langle f_i \rangle \geq f_c$.

| Corpus, | Data summary for relatively common words | | | | | |
|---|---|---|---|---|---|---|
| (1-grams) | $f_c$ | $N_t(words)$ | % (of all words) | $N_{r'}(values)$ | $\langle r' \rangle$ | $\sigma[r']$ |
| English | $5 \times 10^{-8}$ | 106,732 | 1.45 | 16,568,726 | $1.19 \times 10^{-2}$ | 0.98 |
| English fiction | $1 \times 10^{-7}$ | 98,601 | 3.77 | 15,085,368 | $5.64 \times 10^{-3}$ | 0.97 |
| Spanish | $1 \times 10^{-6}$ | 2,763 | 0.124 | 473,302 | $9.00 \times 10^{-3}$ | 0.96 |
| Hebrew | $1 \times 10^{-5}$ | 70 | 0.011 | 6,395 | $3.49 \times 10^{-2}$ | 1.00 |

**Table 13.3**: Summary of *Google* corpus data. Annual growth rates correspond to data in the 209-year period 1800–2008.

| Corpus, | Annual use $u_i(t)$ 1-gram data | | | | | Annual growth $r(t)$ data | | |
|---|---|---|---|---|---|---|---|---|
| (1-grams) | $N_u(uses)$ | $Y_i$ | $Y_f$ | $N_w(words)$ | $Max[u_i(t)]$ | $N_r(values)$ | $\langle r \rangle$ | $\sigma[r]$ |
| English | $3.60 \times 10^{11}$ | 1520 | 2008 | 7,380,256 | 824,591,289 | 310,987,181 | $2.21 \times 10^{-2}$ | 0.98 |
| English fiction | $8.91 \times 10^{10}$ | 1592 | 2009 | 2,612,490 | 271,039,542 | 122,304,632 | $2.32 \times 10^{-2}$ | 1.03 |
| Spanish | $4.51 \times 10^{10}$ | 1532 | 2008 | 2,233,564 | 74,053,477 | 111,333,992 | $7.51 \times 10^{-3}$ | 0.91 |
| Hebrew | $2.85 \times 10^{9}$ | 1539 | 2008 | 645,262 | 5,587,042 | 32,387,825 | $9.11 \times 10^{-3}$ | 0.90 |

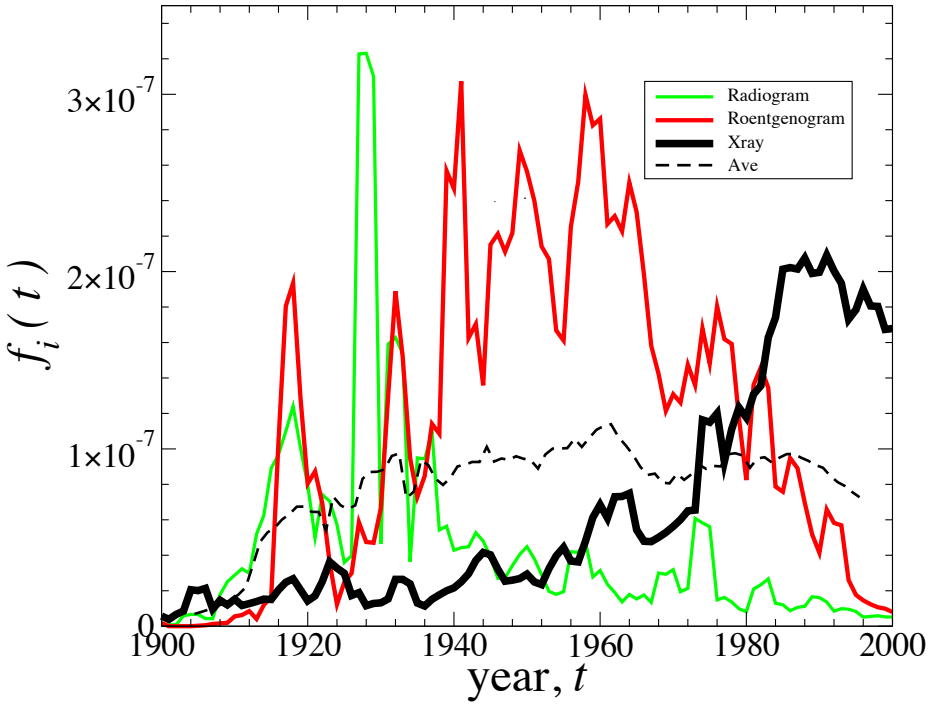


**Figure 13.5**: **Word extinction.** The extinction of the English word "Roentgenogram" as a result of word competition with two competitors, "Xray" and "Radiogram." The average of the three $f_i(t)$ is relatively constant over the 80-year period 1920–2000, indicating that these 3 words were competing for limited linguistic "market share." We conjecture that the higher fitness of "Xray" is due to the efficiency arising from its shorter word length and also due to the fact that English has become the base language for scientific publication.

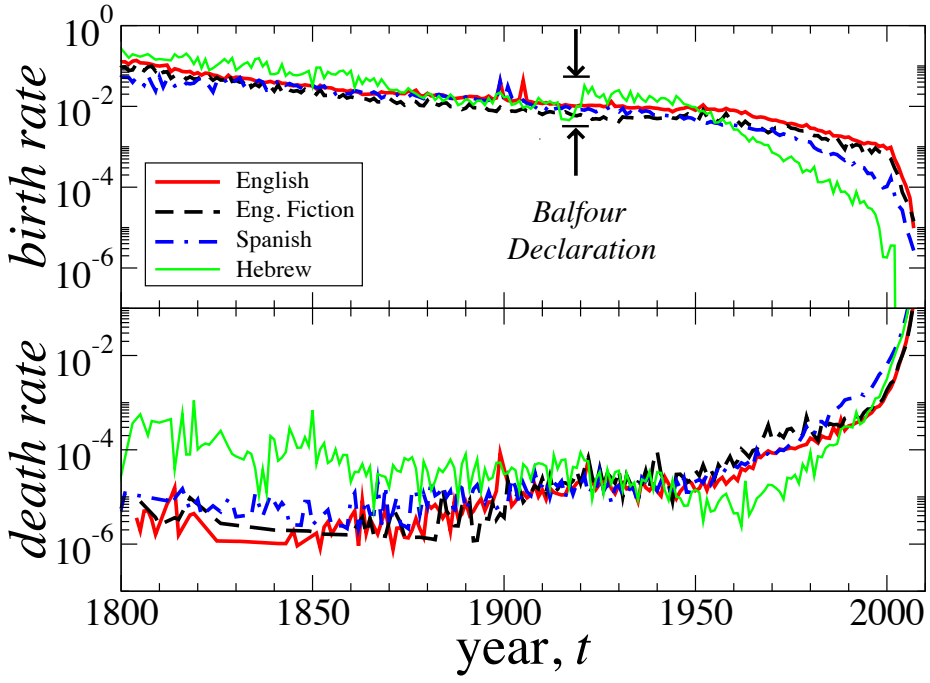


**Figure 13.6**: **Dramatic shift in the birth rate and death rate of words.** The birth rate $r_b$ and the death rate $r_d$ of words demonstrate the inherent time dependence of the competition level between words in each of 4 corpora analyzed. The modern print era shows a marked increase in the death rate of words (e.g. low fitness, misspelled and outdated words). There is also a simultaneous decrease in the birth rate of new words, consistent with the decreasing marginal need for new words. This fact is also reflected by the sub-linear Heaps' law exponent $b < 1$ calculated for all languages in [113]. Note the impact of the Second Aliyah of immigration to Palestine ending in 1914 and the Balfour Declaration of 1917, credited with rejuvenating the Hebrew language as a national language.

.

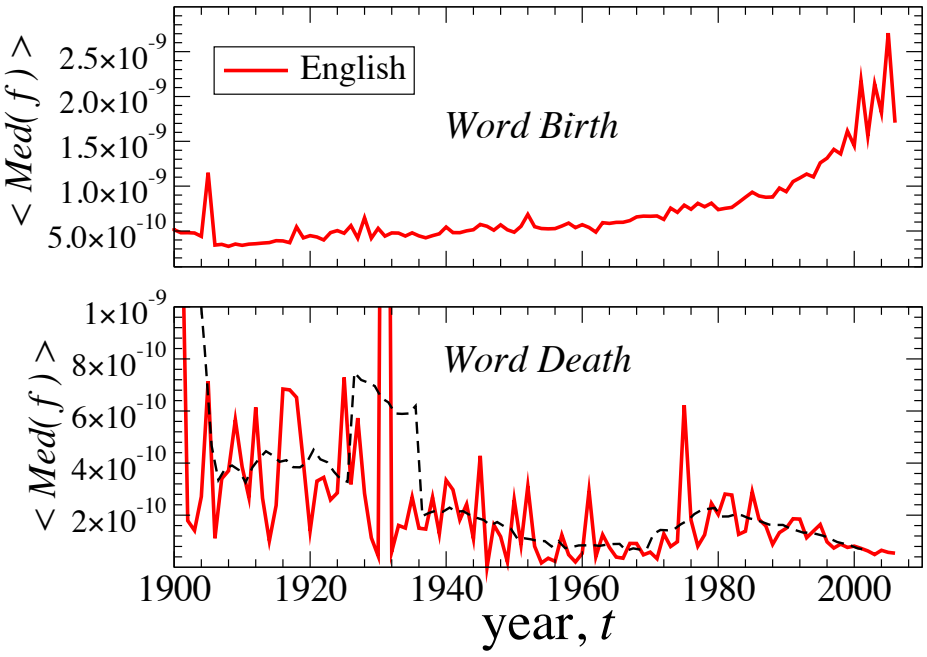


**Figure 13.7**: **Survival of the fittest in the entry process of words.** Trends in the relative uses of words that either were born or died in a given year show that the degree of competition between words is time dependent. For the English corpus, we calculate the average median lifetime relative use $\langle \mathrm{Med}(f_i) \rangle$ for all words $i$ born in year $t$ (top panel) and for all words $i$ that died in year $t$ (bottom panel), which also includes a 5-year moving average (dashed black line). The relative use ("utility") of words that are born shows a dramatic increase in the last 20–30 years, as many new technical terms, which are necessary for the communication of modern devices and ideas, are born with relatively high intrinsic fitness. Conversely, with higher editorial standards and the recent use of word processors which include spelling standardization technology, the words that are dying are those words with low relative use, which we also confirm by visual inspection of the lists of dying words to be misspelled and nonsensical words.

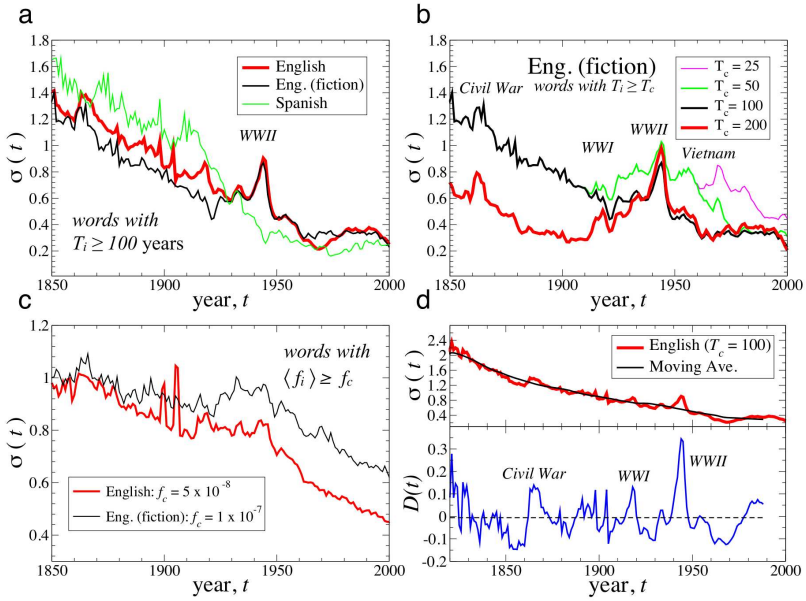


**Figure 13.8**: **Historical events are a major factor in the evolution of word use.** The variation $\sigma(t)$ in the growth rate $r_i(t)$ of relative word use defined in Eq. (12.2) demonstrates the increased variation in growth rates during periods of international crisis (e.g. World War II). The increase in $\sigma(t)$ during the World War II, despite the overall decreasing trend in $\sigma(t)$ over the 159-year period, demonstrates a"globalization" effect, whereby societies are brought together by a common event and a unified media. Such contact between relatively isolated systems necessarily leads to information flow. **(a)** The variation $\sigma(t)$ calculated for the relatively new words with $T_c = 100$. The Spanish corpus does not show an increase in $\sigma(t)$ during World War II, indicative of the relative isolation of South America and Spain from the European conflict. **(b)** $\sigma(t)$ for four sets of post-1800 words $i$ that meet the criteria $T_i \geq T_c$. The oldest "new" words, corresponding to $T_c = 200$, demonstrate the strong increase in $\sigma(t)$ during World War II, with a peak around 1945. **(c)** The standard deviation $\sigma(t)$ in the growth rates $r_i(t)$ for the most common words, defined by words such that $\langle f_i \rangle > f_c$ over the entire lifetime. **(d)** We compare the variation $\sigma(t)$ for common words with the 20-year moving average over the time period 1820–1988, which also demonstrates an increasing $\sigma(t)$ during times of national/international crisis, such as the American Civil War (1861–1865), World War I (1914–1918) and World War II (1939–1945), and recently during the 1980s and 1990s, possibly as a result of new digital media which offer new environments for the dynamics of word use. $D(t)$ is the difference between the moving average and $\sigma(t)$.

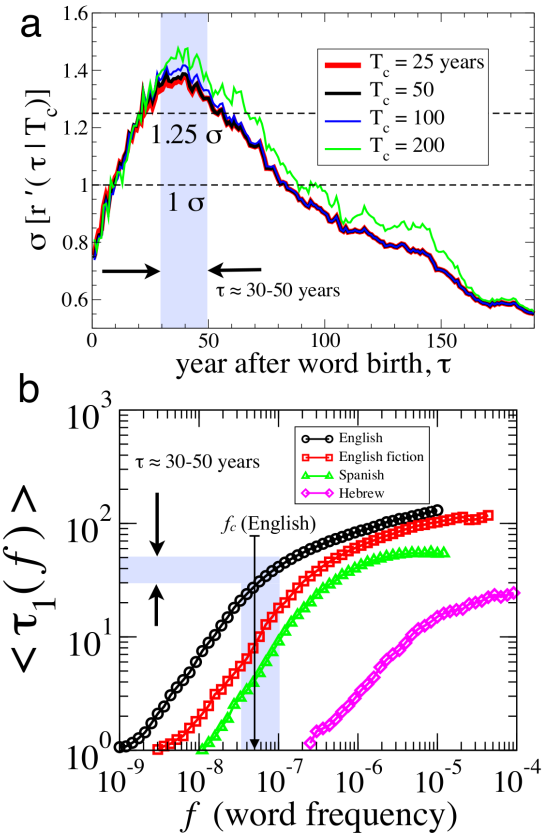


**Figure 13.9**: **Quantifying the tipping point for word use. (a)** The maximum in the standard deviation $\sigma$ of growth rates during the "adolescent" period $\tau \approx 30$–$50$ indicates the characteristic time scale for words being incorporated into the standard lexicon, i.e. inclusion in popular dictionaries. In Fig. 13.4 we plot the average growth rate trajectory $\langle r'(\tau|T_c)\rangle$ which also shows relatively large positive growth rates during approximately the same period $\tau \approx 30$–$50$ years. **(b)** The first passage time $\tau_1$ [143] is defined as the number of years for the relative use of a new word $i$ to exceed for the first time a given $f$-value, defined here by the first instance in the corpus that the a given word $i$ satisfies $f_i[\tau_1(f)] \geq f$, can also be used to quantify the thresholds for sustainability for new words. The average first-passage time $\langle\tau_1(f)\rangle$ to $f_c \equiv 5 \times 10^{-8}$ for the English corpus, (recall $f_c$ represents the threshold for a word belonging to the "kernel" lexicon), roughly corresponds to the peak time $\tau \approx 30 - 50$ years in $\sigma(\tau)$ shown in panel (a). This feature supports our conjecture that the peak in $\sigma(\tau)$ reflects the time scale over which a word is accepted into the standard lexicon.

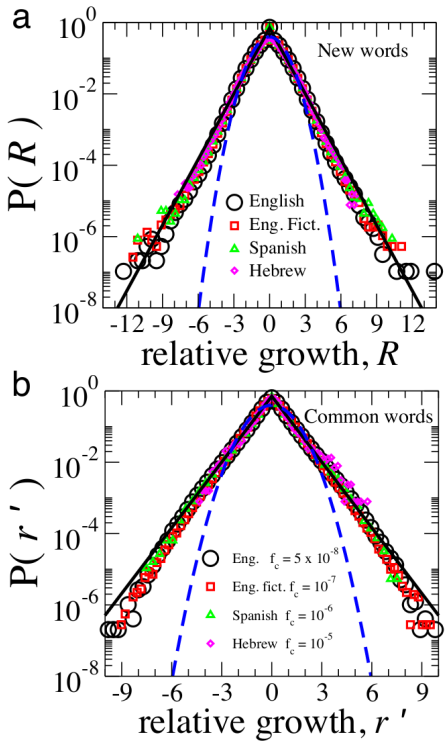


**Figure 13.10**: **Common growth distribution for new words and common words. (a)** We find Laplace distributions, defined in Eq. (13.2), for the annual word use growth rates for relatively new words, as well as for relatively common words for English, Spanish and Hebrew. These growth distributions, which are symmetric and centered around $R \approx 0$, exhibit an excess number of large positive and negative values when compared with the Gaussian distribution. The Gaussian distribution (dashed blue) is the predicted distribution for the Gibrat growth model.[117]. We analyze word use data over the time period 1800-2008 for new words $i$ with lifetimes $T_i \geq 100$ years years (see SI methods section and Table 13.1 for a detailed description). **(b)** PDF $P(r')$ of the annual relative growth rate $r'$ for dataset ii words which have average relative use $\langle f_i \rangle \geq f_c$. These select words are relatively common words. In order to select relatively frequently used words, we use the following criteria: $T_i \geq 10$ years, $1800 \leq t \leq 2008$, and $\langle f_i \rangle \geq f_c$. There is no need to account for the age-dependent trajectory $\sigma[r'(\tau|T_c)]$, as in the normalized growth defined in Eq. (15.5), for these relatively common words since they are all most likely in the mature phase of their lifetime trajectory.

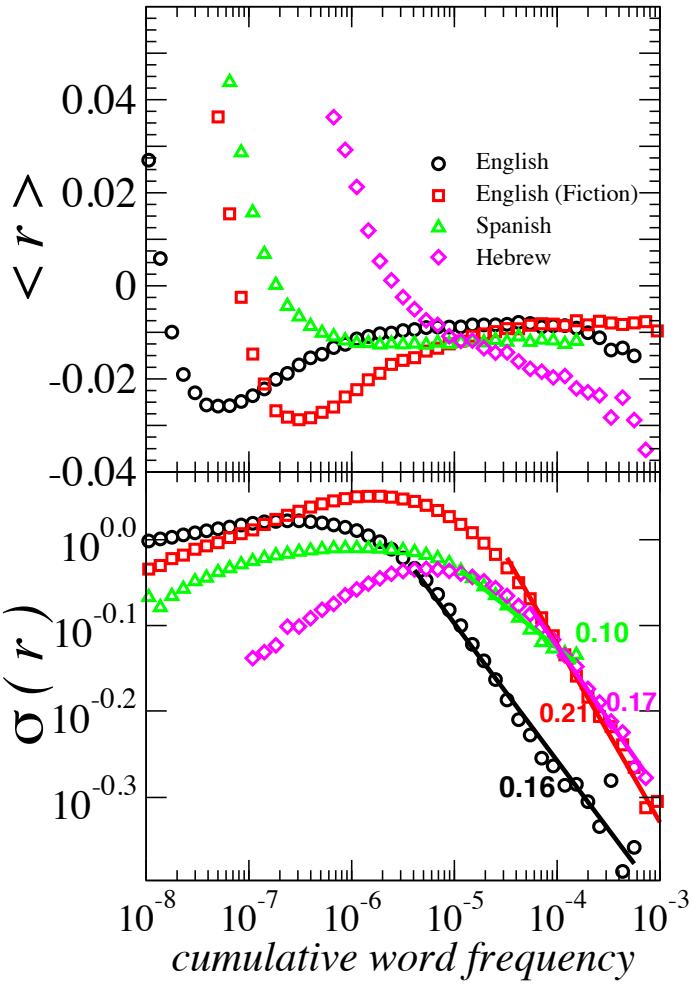


**Figure 13.11**: **Scaling in the "volatility" of common words.** The dependence of growth rates on the cumulative word frequency $S_i \equiv \sum_{t'=0}^{t} f_i(t)$ calculated for a combination of new [dataset (i)] and common [dataset (ii)] words that satisfy the criteria $T_i \geq 10$ years (similar results for threshold values $T_c = 50, 100$, and 200 years). **(a)** Average growth rate $\langle r \rangle$ saturates at relatively constant (negative) values for large $S$. The negative values may represent a "crowding out" effect in taking place in a dynamic corpus. **(b)** Scaling in the standard deviation of growth rates $\sigma(r|S) \sim S^{-\beta}$ for words with large $S$, also observed for the growth rates of large economic institutions [119, 121]. Here this size-variance relation corresponds to scaling exponent values $0.10 < \beta < 0.21$, which are related to the non-trivial bursting patterns and non-trivial correlation patterns in literature topicality. We calculate $\beta_{Eng.} \approx 0.16 \pm 0.01$, $\beta_{Eng.fict} \approx 0.21 \pm 0.01$, $\beta_{Spa.} \approx 0.10 \pm 0.01$ and $\beta_{Heb.} \approx 0.17 \pm 0.01$.

# Chapter 14

# Discussion

The digital era has brought forth a data deluge making possible empirical studies of almost every aspect of human activity [128]. The digitization of written language has resulted in a rapid increase in cultural trend analysis using word frequencies and other quantities extracted from big data sources using natural language processing methods. The amount of metadata extractable from internet feeds is dizzying and subject to various definitions of relevance. For example, written language published in blogs on the daily timescale can be vaguely categorized into "obscure blogs", "more popular blogs", "tech columns", and "mainstream news coverage." As a result of this coarse hierarchical schema, analysis of online corpora is subject to significance thresholds that can be difficult to define. However, there are well-defined entry requirements for published documents, which must meet editorial standards and conform to the principles of market supply-demand.

Despite the careful guard of libraries around the world which house the written corpora for almost every written language, little is known about the aggregate dynamics of word evolution over the characteristic time scale of a human generation. Before the digitization of written language, the analysis of social and political trends required painstaking brute force manual work and rules of thumb so that a quantitative analysis of cultural trends would suffer from small sample effects since quantitative measurements would miss the large number of topics at any given point that are below the threshold for detection, and long-term analysis would suffer from

repeatedly crossings of topics above and below reliable significance thresholds due to intrinsic fluctuations. However, the massive Google Books database allows social scientists to make reliable studies of word evolution at the ecosystem level which is far beyond the level of individual word case studies.

Inspired by similar research on the growth of business firms and subsequent extensions to a wide range of competition driven systems - from countries and bird populations to religious activity and scientific journals - we extend the concepts and methods to the dynamics of word evolution. Might words be understood as competing actors in a system of finite resources? Just as business firms compete for market share and scientific journal compete for reader attention, could words demonstrate the same growth statistics because they are competing for the use of the writer/speaker and competes for the attention of the corresponding reader/listener [129, 130, 131, 111, 132]?

Indeed, we observe striking similarity between the distribution of growth rates for words in the Google Books data and what was found in the various competing dynamics mentioned above. We also document the case example of Xray (Fig. 1) which suggests that a finite number of categorically related words will compete in a zero-sum game. Further, such activity does not take place in a vacuum. We find that the dynamics in this competitive arena are influenced by historical context, trends in global communication, and the means for standardizing that communication. Just as there are recessions and booms in a global economy, the marketplace for words waxes and wanes with a global pulse as historical events unfold. And just as regulators put limits on economic risk and market domination, standardization technologies such as the dictionary and spell checkers serve as powerful arbiters in determining the characteristic pace of word evolution. Since the context of word use is so important, we anticipate that niches [125] in various language ecosystems (ranging from spoken word to professionally published documents to various online forms such as chats, tweets and blogs) have heterogeneous selection laws that may favor a given word in one arena but not another. Moreover, the birth and death rate of words and their close associates (misspellings, synonyms, abbreviations) also likely depend on factors endogenous (internal) to the language domain such as correlations in word use to other partner words and polysemous contexts [106, 107] as well as exogenous (external) socio-technological factors and demographic aspects of the writers, such as

age [107] and social niche [125].

One intrinsic timescale in evolutionary systems is the reproduction age of the interacting hosts. Interestingly, we find a pronounced peak in the fluctuations of word growth rates when a word has reached approximately 30-50 years of age, see Fig. 13.9(a), and posit that this the timescale for a word to be accepted into a standardized dictionary which inducts words that are used above a threshold frequency, consistent with the first-passage times to $f_c$ in Fig. 13.9(b), which is corroborated by the related frequencies associated with standardized dictionaries compared in Michel et al. [100]. This timescale roughly corresponding to a characteristic human lifetime scale, and points to the generational features of cultural transmission as a strong factor in the evolution of language. The prominent role of new generation of speakers in language evolution has precedent in linguistics. For example, it has been shown that primitive pidgin languages, which are little more than crude mixes of parent languages, spontaneously acquire the full range of complex syntax and grammar once they are learned by the children of a community as a native language. It is at this point a pidgin becomes a creole, in a process referred to as nativization [109].

Nativization also had a prominent effect in the revival of the Hebrew language, a significant historical event which also manifests prominently in our statistical analysis. The birth rate of new words in the Hebrew language jumped by a factor of 5 in just a few short years around 1920 following the Second Aliyah wave of immigration to Israel and the Balfour Declaration of 1917. The combination of new communities of like-minded young people, living in relatively closed-off social cells speaking Hebrew together with the Balfour Declaration, in which the British government explicitly endorsed the establishment of a national homeland for the Jewish people in the Palestine Mandate, resulted in the group consensus along two issues: (i) that the Hebrew language, hitherto used largely only for writing, was gaining official status as a modern spoken language, and (ii) that a national community of Jews along with a centralized culture would become an increasingly plausible notion. The unique history of the Hebrew language in concert with the Google Books data thus provide an unprecedented opportunity to quantitatively study the emerging dynamics of what is in some regards a new language.

The impact of historical context on language dynamics is not limited to emerg-

ing languages, however, but extends to languages that have been active and evolving continuously for a thousand years. We find that historical episodes perturb the competitive arena of existing languages much like exogenous shocks can perturb the stock market [133, 114, 115]. Specifically, we find that the distribution of word growth rates broadens markedly during times of large scale conflict, such as World War II. This can understood as manifesting from the unification of public consciousness that creates fertile breeding ground for new ideas. People are less likely to have their attention drawn to local events and the language of business as usual and more likely to be focused on the overwhelming impact of the current global conflict. Remarkably, these effects do not leak over into unaffected regions and their associated languages, as we observe no such broadening in the Spanish language corpus during World War II, even as the war's impact on English word selection is unequivocal. As most Spanish speaking countries had minor if not absentee roles in WWII, it's not surprising that there was no endemic contribution from countries to the war's linguistic effects, but it is notable that the activity from one language should not show any "leakage" into another. However, this phenomenon too is not without analogous examples, as a large number of the world's ethnic groups are separated along linguistic lines, showing just how effective a language barrier is in isolating populations.

This study is motivated by analogies with other competition driven systems systems, such as the growth dynamics of companies [116, 117, 118, 134, 135, 136], countries [119, 117, 137], universities [138], journals [139], religious activities [140], careers [141] and animal populations [142]. We find a striking analogy between the relative use of a word, which can quantitatively represent the intrinsic value of the word, and the value of a company (e.g. measured by its market capitalization or sales). This suggests a common underlying mechanism: just as firms compete for market share leading to business opportunities, and animals compete for food and shelter leading to reproduction opportunities, words are competing for use among the books that constitute a corpus.

# Chapter 15

# Methods

**Quantifying the word use trajectory.** Next we ask how word use evolves through the various stages of its lifetime. Since words appear to compete for use in a word-space that is based on utility, we seek to quantify the average lifetime trajectory of word use. The lifetime trajectories of different words will vary, since each trajectory depends not only on the intrinsic utility of word $i$, but also on the "birth-year" $t_{0,i}$ of word $i$.

Here we define the age or trajectory year $\tau = t - t_{0,i}$ as the number of years after the word's first appearance in the database. In order to compare word use trajectories across time and across varying utility, we normalize the trajectories for each word $i$ by the average use

$$\langle f_i \rangle \equiv \frac{1}{T_i} \sum_{t=t_{0,i}}^{t_{f,i}} f_i(t) \tag{15.1}$$

over the lifetime $T_i \equiv t_{f,i} - t_{0,i} + 1$ of the word, leading to the normalized trajectory,

$$f_i'(\tau) = f_i'(t - t_{i,0}|t_{i,0}, T_i) \equiv f_i(t - t_{i,0})/\langle f_i \rangle \ . \tag{15.2}$$

By analogy, in order to compare various growth trajectories, we normalize the relative growth rate trajectory $r_i'(t)$ by the standard deviation over the entire lifetime,

$$\sigma[r_i] \equiv \sqrt{\frac{1}{T_i} \sum_{t=t_{0,i}}^{t_{f,i}} [r_i(t) - \langle r_i \rangle]^2} \ . \tag{15.3}$$

Hence, the normalized relative growth trajectory is

$$r'_i(\tau) = r'_i(t - t_{i,0}|t_{i,0}, T_i) \equiv r_i(t - t_{i,0})/\sigma[r_i] \ . \tag{15.4}$$

Using these normalized trajectories, Fig. 13.4 shows the weighted averages $\langle f'(\tau|T_c)\rangle$ and $\langle r'(\tau|T_c)\rangle$ and the weighted standard deviations $\sigma[f'(\tau|T_c)]$ and $\sigma[r'(\tau|T_c)]$. We compute $\langle\cdots\rangle$ and $\sigma[\cdots]$ for each trajectory year $\tau$ using all $N_t$ trajectories (Table 13.1) and using all words that satisfy the criteria $T_i \geq T_c$ and $t_{i,0} \geq 1800$. We compute the weighted average and the weighted standard deviation using $\langle f_i\rangle$ as the weight value for word $i$, so that $\langle\cdots\rangle$ and $\sigma[\cdots]$ reflect the lifetime trajectories of the more common words that are "new" to each corpus.

We analyze the relative growth of word use in a fashion parallel to the economic growth of financial institutions, and show in Fig. 13.10(b) that the pdf $P(r')$ for the relative growth rates is not only centered around zero change corresponding to $r \approx 0$ but is also symmetric around this average. Hence, for every word that is declining, there is another word that is gaining by the same relative amount. Since there is an intrinsic word maturity $\sigma[r'(\tau|T_c)]$ that is not accounted for in the quantity $r'_i(\tau)$, we further define the detrended relative growth

$$R \equiv r'_i(\tau)/\sigma[r'(\tau|T_c)] \tag{15.5}$$

which allows us to compare the growth factors for new words at various life stages. The result of this normalization is to rescale the standard deviations for a given trajectory year $\tau$ to unity for all values of $r'_i(\tau)$. Fig. 13.10 shows common growth patterns $P(R)$ and $P(r')$, independent of corpus. Moreover, we find that the Laplace distributions $P(R)$ found for the growth rates of word use are surprisingly similar to the distributions of growth rates for economic institutions of varying size, such as scientific journals, small and large companies, universities, religious institutions, entire countries and even bird populations [119, 137, 136, 118, 116, 138, 134, 117, 139, 135, 141, 142].

**Quantifying the long-term social memory.** In order to gain understanding of the overall dynamics of word use, we have focused much of our analysis on the distributions of $f_i$ and $r_i$. However, distributions of single observation values discard information about temporal ordering. Hence, in this section we also examine the

temporal correlations in each time series $f_i(t)$ to uncover memory patterns in the word use dynamics. To this end, we compare the autocorrelation properties of each $f_i(t)$ to the well-known properties of the time series corresponding to a 1-dimensional random walk.

In a time interval $\delta t$, a time series $Y(t)$ deviates from the previous value $Y(t-\delta t)$ by an amount $\delta Y(t) \equiv Y(t) - Y(t-\delta t)$. A powerful result of the central limit theorem, also known as Fick's law of diffusion, is that if the displacements are independent (uncorrelated corresponding to a simple Markov process), then the total displacement $\Delta Y(t) = Y(t) - Y(0)$ from the initial location $Y(0) \equiv 0$ scales according to the total time $t$ as

$$\Delta Y(t) \equiv Y(t) \sim t^{1/2} \ . \tag{15.6}$$

However, if there are long-term correlations in the time series $Y(t)$, then the relation is generalized to

$$\Delta Y(t) \sim t^{H} \ , \tag{15.7}$$

where $H$ is the Hurst exponent which corresponds to positive correlations for $H > 1/2$ and negative correlations for $H < 1/2$.

Since there may be underlying social, political, and technological trends that influence each time series $f_i(t)$, we use the detrended fluctuation analysis (DFA) method [123, 124, 122] to analyze the residual fluctuations $\Delta f_i(t)$ after we remove the local linear trends using time windows of varying length $\Delta t$. The time series $\tilde{f}_i(t|\Delta t)$ corresponds to the locally detrended time series using window size $\Delta t$. Hence, we calculate the Hurst exponent $H$ using the relation between the root-mean-square displacement $F(\Delta t)$ and the window size $\Delta t$ [123, 124, 122],

$$F(\Delta t) = \sqrt{\langle \Delta \tilde{f}_i(t|\Delta t)^2 \rangle} = \Delta t^{H} \ . \tag{15.8}$$

Here $\Delta \tilde{f}_i(t|\Delta t)$ is the local deviation from the average trend, analogous to $\Delta Y(t)$ defined above.

Fig. 13.2 shows 4 different $f_i(t)$ in panel (a), and plots the corresponding $F_i(\Delta t)$ in panel (b). The calculated $H_i$ values for these 4 words are all significantly greater than the uncorrelated $H = 0.5$ value, indicating strong positive long-term correlations in the use of these words, even after we have removed the local trends. In these cases, the trends are related to political events such as war in the cases of "Americanism"

and "Repatriation", or the bursting associated with new technology in the case of "Antibiotics," or new musical trends in the case of "polyphony."

In Fig. 13.3 we plot the pdf of $H_i$ values calculated for the relatively common words analyzed in Fig. 13.10(b). We also plot the pdf of $H_i$ values calculated from shuffled time series, and these values are centered around $\langle H \rangle \approx 0.5$ as expected from the removal of the intrinsic temporal ordering. Thus, using this method, we are able to quantify the social memory characterized by the Hurst exponent which is related to the bursting properties of linguistic trends, and in general, to bursting phenomena in human dynamics [114, 115, 126, 127].

# Part VI

# Conclusion

# Chapter 16

# Conclusion

This thesis covers work done in three distinct systems where the complex emergent phenomena are fundamentally related to the large number of individual components, interacting at various scales, often with a certain degree of internal hierarchy. Drawing on methods and concepts from statistical physics, we search for statistical patterns that emerge from the complex interactions between components in three distinct settings: (i) earthquakes, (ii) financial systems, and (iii) human use of language.

The impact and applications for earthquake research are easily the most accessible. As demonstrated by the tragic consequences of the Great East Japan Earthquake of March 2011, the resulting tsunami and nuclear accidents of which making it the most expensive natural disaster in human history, important improvements can still be made in earthquake risk management today, more a century after seismology first emerged as a field of study. While knowledge of plate boundaries, fault locations, and slip rates are all improving, such knowledge is still sparse and by its nature difficult to know with high precision and accuracy. Even given these, major earthquakes can still evade mechanistic prediction. Importing the concept of network analysis, our work shows additional factors to consider in connecting the great locational chain of earthquake interdependence. While earthquakes may remain near impossible to predict, improved risk analysis facilitated by our work may make for more judicious planning in high-risk areas in the future.

Financial models and their application to both financial and nonfinancial systems

too have implications on future research and knowledge. In the last century, economics became an empirical science with the advent of macroscopic measures like country GDP, and microscopic measures like the price of a stock on a market, which economists have introduced "hard" quantitative modeling on through the use of econometrics. At a high level, the physicists' perspective, which includes concepts such as scaling, universality, stationarity, symmetry, and random walk diffusion, may aide in drawing helpful connections between otherwise unconnected observations. At a ground level, physicists are free report unexplained, but nonetheless interesting observations while economists are more cautious, often not reporting what cannot be explained by a new formal model, commonly an analytically soluble one. Finally, the quantitative lens of a statistical physicist has by its nature and history had a greater flexibility in being applied to ostensibly unrelated disciplines like linguistics. Methods in physics may prove a valuable supplement to the more conventional means of research in various areas.

Of interest to anyone living in a modern interconnected society is the question of what a large market crash looks like. Evidence that large crashes aren't yet adequately described is easily found in the form of the post-2008 world recession. According to standard theories in economics, this crash and others like it essentially shouldn't exist. In terms of conventional description, the odds are simply too low, yet large crashes typically occur once a decade. With this motivation in mind, a simple question to ask is, "Are crashes universal in nature across countries?" A physics law observed in one location presumably holds anywhere in the world and a physics law observed in one system will hold in every other system of that same category. Extending beyond spatial universality, we can also ask if characteristics of crashes are constant across time, as many conserved quantities in physics are.

Finally, physics can help make sense of large sets of social and linguistic data hitherto unavailable. The modern era is marked by an unprecedented ability to quantify human behavior as it relates to everyday life. More and more, the limit of understanding is not held back by dearth of data, but by the inability to make headway through its over-lavish abundance. There is so much information in front of us, we don't know where to start. Here too, the concepts and tools of statistical physics can provide an intuitive starting point and a guiding compass. The traditional

academic fields for studying these concepts (e.g. linguistics, sociology) have relied qualitative rules-of-thumb and painstakingly eking out patterns while physics in many cases has the power to draw emergent trends out of the aggregate. A linguist may make note of when a particular word has passed his/her threshold for detection as to have joined mainstream usage, but a statistical physicist can infer exactly how long the majority of words take to reach the tipping point of popularity by observing fluctuations and first passage times.

Of course, not everything that can be thought of is a good idea. Not all concepts will map one-to-one from physics to related fields and there runs a risk of blindly overinterpreting results from a field one is not trained in. But, paraphrasing George E. P. Box, while all models may be wrong, some can prove useful. Given restraint and due deference to the existing knowledge in the respective fields and a critical eye for applicability, methods of physics offer an elegant complement to traditional studies. The combination can easily be more quantitative, hence more actionable, and, at a deeper level, more philosophically satisfying.

# References

# Bibliography


[1] B. Gutenberg and C. F. Richter, *Bulletin of the Seismological Society of America* **34**, 185 (1944).

[2] F. Omori, *Journal of the College of Science*, Imperial University of Tokyo **7**, 111-200 (1894); see the recent work of M. Bottiglieri, L. de Arcangelis, C. Godano, and E. Lippiello, *Physical Review Letters* **104**, 158501 (2010).

[3] M. Bath, *Tectonophysics*, **2**, 483 (1965).

[4] Y. Y. Kagan and L. Knopoff, *Geophysical Journal of the Royal Astronomical Society* **62**, 303 (1980).

[5] D. Turcotte, *Fractals and Chaos in Geology and Geophysics* (Cambridge University Press, Cambridge, England, 1997).

[6] P. Bak, K. Christensen, L. Danon, and T. Scanlon, *Physical Review Letters* **88**, 178501 (2002).

[7] A. Corral, *Physical Review E* **68**, 035102(R) (2003).

[8] R. Olsson, *Geodynamics* **27**, 547 (1999).

[9] S. Abe and N. Suzuki, *European Physical Journal B* **59**, 9397 (2007).

[10] S. Abe and N. Suzuki, *Physica A* **332**, 533 (2004).

[11] S. Abe and N. Suzuki, *Journal of Geophysical Research* **108**, 2113 (2003).

[12] S. Abe and N. Suzuki, *Physica A* **350**, 588 (2005).

[13] J. Davidsen, P. Grassberger, and M. Paczuski, *Physical Review E* **77**, 066104 (2008)

[14] M. Baiesi and M. Paczuski, *Physical Review E* **69**, 066106 (2004).

[15] V. N. Livina, S. Havlin, and A. Bunde, *Physical Review Letters* **95**, 208501 (2005).

[16] E. Lippiello, L. de Arcangelis, and C. Godano, *Physical Review Letters* **100**, 038501 (2008).

[17] S. Lennartz, V. N. Livina, A Bunde, and S. Havlin, *Europhysics Letters* **81**, 69001 (2008).

[18] A. Corral, *Physical Review Letters* **92**, 108501 (2004).

[19] To detrend the data, we obtain a best fit linear trend for each time series and subtract it from the series. We calculate the cross-correlation between the detrended sequences.

[20] J. Davidsen and C. Goltz, *Geophysical Research Letters* **31** L21612 (2004).

[21] K. R. Felzer, T. W. Becker, R. E. Abercrombie, G Ekstrom, and J. R. Rice, *Journal of Geophysical Research* **107**, 2190 (2002).

[22] J. L. Hardebeck, K. Felzer, and A. J. Michael, *Journal of Geophysical Research* **113**, B08310 (2008).

[23] J. B. Bassingthwaighte, L. S. Liebovitch, and B. J. West, *Fractal Physiology* (Oxford U. Press, New York, 1994).

[24] R. F. Engle, *Econometrica* **50**, 987 (1982).

[25] Y. Ashkenazy, P. Ch. Ivanov, S. Havlin, C.-K. Peng, A. L. Goldberger, and H. E. Stanley, *Physical Review Letters* **86**, 1900-1903 (2001).

[26] A. Corral, *Physical Review Letters* **95**, 159801 (2005).

[27] E. Lippiello, L. de Arcangelis, and C. Godano, *Physical Review Letters* **100**, 038501 (2008).

[28] K. Pearson, *Nature* **72**, 294 (1905).

[29] R. F. Engle and V. Ng, *Journal of Finance* **48**, 1749 (1993).

[30] P. C. Wason. "The Processing of Positive and Negative Information," *The Quarterly Journal of Experimental Psychology*, **11** (2), 92107 (1959)

[31] R. Baumeister, E. Bratslavsky, C. Finkenauer, and K. Vohs. "Bad is Stronger than Good," *Review of General Psychology* **5** (4): 323370 (2001) is considered a definitive publication on the topic.

[32] C. W. J. Granger and Z. Ding, *Journal of Econometrics* **73**, 61 (1996). Their observation of a long-range correlation in volatility was made more quantitative in the extensive analysis of P. Cizeau, Y. Liu, M. Meyer,C. K. Peng,H. E. Stanley, *Physica A* **245**, 441-445 (1997) and Y. Liu, P. Gopikrishnan, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, *Physical Review E* **60**, 1390-1400 (1999).

[33] B. Podobnik, P. Ch. Ivanov, K. Biljakovic, D. Horvatic, H. E. Stanley, and I. Grosse, *Physical Review E* **72**, 026121 (2005).

[34] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, *Physical Review E* **49**, 1685-1689 (1994).

[35] J. D. Hamilton, *Time Series Analysis* (Princeton, New Jersey, 1994).

[36] `www.physionet.org.pn3.ucddb/`

[37] The subjects are selected for possible diagnosis of obstructive sleep apnea or primary snoring. The subject's details are available at tt www.physionet.org.pn3.ucddb/SubjectsDetails.xls. We use time series designated as uccdbb0i.eeg2, where $i$ stands for an integer.

[38] C. P.Pan, B. Zheng, Y. Z. Wu, Y. Wang and X. W. Tang , *Physics Letters A* **329**, 130 (2004).

[39] B. Podobnik and H. E. Stanley, *Physical Review Letters* **100**, 084102 (2008).

[40] E. F. Fama, *The Journal of Business* **38**, 285 (1965).

[41] C. Hiemstra and J. D. Jones, *Journal of Empirical Finance* **4**, 373 (1997).

[42] J. T. Barkoulas and C. F. Baum, *Economics Letters* **53**, 253 (1996).

[43] T. Jagric, B. Podobnik, and M. Kolanovic, *Eastern European Economics* **43**, 79 (2005).

[44] B. Podobnik, D. Fu, T. Jagric, I. Grosse, and H. E. Stanley, *Physica A* **362**, 465 (2006).

[45] K. Matia, M. Pal, H. Salunkay, and H. E. Stanley, *Europhysics Letters* **66**, 909 (2004).

[46] T. Lux, *Applied Financial Economics* **6**, 463 (1996).

[47] P. Gopikrishnan, M. Meyer, L. A. N. Amaral, and H. E. Stanley, *European Physical Journal B* **3**, 139 (1998),

[48] P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, *Physical Review E* **60**, 5305 (1999).

[49] V. Plerou, P. Gopikrishnan, L. A. N. Amaral, M. Meyer, and H. E. Stanley, *Physical Review E* **60**, 6519 (1999).

[50] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley, *Proceedings of the National Academy of Sciences* **106**, 22079 (2009).

[51] U. A. Muller, M. M. Dacorogna, R. B. Olsen, O. V. Pictet, M. Schwarz, and C. Morgenegg, *Banking Finance* **14** 1189 (1995).

[52] M. M. Dacorogna, U. A. Muller, R. J. Nagler, R. B. Olsen, and O. V. Pictet, *Journal of International Money and Finance* **12** 413 (1993).

[53] H. E. Hurst,*Proceedings of the Institution of Civil Engineers* **1**, 519 (1951).

[54] M. Casandro and G. Jona-Lasinio, *Advances in Physics* **27**, 913 (1978).

[55] H. E. Stanley and N. Ostrowsky, eds.: *Correlations and Connectivity: Geometric Aspects of Physics, Chemisty and Biology* (Kluwer, Dordrecht, 1990).

[56] T. Vicsek, *Fractal Growth Phenomenon, 2nd Edition* (World Scientific, Singapore, 1993).

[57] A. Bunde and S. Havlin (Editors), *Fractals in Science* (Springer, Berlin, 1994).

*Fractal Physiology* (Oxford U. Press, New York, 1994).

[58] J. Beran, *Statistics for Long-Memory Processes* (Chapman & Hall, New York, 1994).

[59] H. Takayasu, *Fractals in the Physical Sciences* (Manchaster U. Press, Manchaster, 1997).

[60] C. W. J. Granger and R. Joyeux, *Journal of Time Series Analysis* **1**, 15 (1980).

[61] J. Hosking, *Biometrika* **68**, 165 (1981).

[62] C. K. Peng *et al.*, *Physical Review A* **44**, 2239(R) (1991).

[63] H. A. Makse *et al.*, *Physical Review E* **53**, 5445 (1995).

[64] L. Hui and E. Gaztanaga, *The Astrophysical Journal* **519**, 622 (1999).

[65] International Human Genome Sequence Consortium,*Nature* **409**, 860 (2001); J. C. Venter *et al.*, *Science* **291**, 1304 (2001).

[66] M. Samon and F. Curley, *Journal of Applied Physiology* **83**, 975 (1997).

[67] C. L. Ehlers, J. W. Havstad,D. Prichard, and J. Theiler, *Journal of Neuroscience* **18**, 7474 (1998).

[68] C. Braun, P. Kowallik, A. Freking, D. Hadeler, K. -D. Kniffki, and M. Meesmann, *American Journal of Physiology - Heart and Circulatory Physiology* **275**, H1577 (1998).

[69] G. Boffetta, A. Celani, and M. Vergassola, *Physical Review E* **61**, 29(R) (2000).

[70] P. Silvapulle and C. W. J. Granger, *Quantitative Finance* **1**, 542 (2001).

[71] A. Ang and J. Chen, *Journal of Financial Economics* **63**, 443 (2002).

[72] K. Ohashi, L. A. N. Amaral, B. H. Natelson, and Y. Yamamoto, *Physical Review E* **68**, 065204(R) (2003).

[73] B. Podobnik, P. Ch. Ivanov, V. Jazbinsek, Z. Trontelj, H. E. Stanley, and I. Grosse, *Physical Review E Rapid Communication* **71**, 025104 (2005).

[74] T. Bollerslev, *Journal of Econometrics* **31**, 307 (1986).

[75] L. Glosten, R. Jagannathan, and D. Runkle, *Journal of Empirical Finance* **48**, 1779 (1993).

[76] R. F. Engle and V. Ding, *Journal of Empirical Finance* **48**, 1749 (1993).

[77] J. M. Zakoian, *Journal of Economic Dynamics and Control* **18**, 931 (1994).

[78] R. F. Engle and A. J. Patton, *Quantitative Finance* **1**, 237 (2001).

[79] E. Sentana, *Review of Economic Studies* **62**, 639 (1995).

[80] R. Rossiter and S. A. Jayasuriya, *Journal of International Finance and Economics* **8**, 11 (2008).

[81] J. H. Wright, "Long memory in emerging market stock returns", *Emerging Markets Quarterly* **5**,50 (2001) pp. 5055.

[82] E. Panas, *Applied Financial Economics* **38**, 395 (2001).

[83] T. Jagric, B. Podobnik, and M. Kolanovic, *Eastern European Economics* **43**, 79 (2005).

[84] Gretl package can be taken from http://packages.debian.org/search?keywords=gretl

[85] J. Campell and L. Hentschel, *Journal of Financial Economics* **31**, 281 (1992).

[86] G. Wu, *Review of Financial Studies* **14**, 837 (2001).

[87] T. Ane, L. Ureche-Rangau, J. B. Gambet, and J. Bouverot, *Journal of International Financial Markets, Institutions and Money* **18**, 326 (2008).

[88] A. M. Petersen, W-S. Jung, J-S. Yang, H. E. Stanley, "Quantitative and Empirical demonstration of the Matthew Effect in a study of Career Longevity," *Proceedings of the National Academy of Sciences* **108**, 18-23 (2011).

[89] A. M. Petersen, O. Penner, H. E. Stanley, "Methods for detrending success metrics to account for inflationary and deflationary factors," *European Physical Journal B* **79**, 67-78 (2011).

[90] J. A. Davies, "The individual success of musicians, like that of physicists, follows a stretched exponential distribution," *European Physical Journal B* **27**, 445-447 (2002)

[91] A. Chatterjee, S. Sinha, B. K. Chakrabarti, "Economic inequality: is it natural?" *Current Science*, **92** (10). pp. 1383-1389. ISSN 0011-3891 (2007)

[92] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA 1949).

[93] A. A. Tsonis, C. Schultz, P. A. Tsonis, "Zipf's law and the structure and evolution of languages," *Complexity* **3**, 12–13 (1997).

[94] M.Á. Serrano, A. Flammini, F. Menczer, "Modeling Statistical Properties of Written Text," *PLoS ONE* **4(4)**, e5372 (2009).

[95] R. Ferrer i Cancho, R. V. Solé, "Two regimes in the frequency of words and the origin of complex lexicons: Zipf's law revisited," *Journal of Quantitative Linguistics* **8**, 165–173 (2001).

[96] R. Ferrer i Cancho, "The variation of Zipfś law in human language," *European Physical Journal B* **44**, 249–257 (2005).

[97] R. Ferrer i Cancho, R. V. Solé, "Least effort and the origins of scaling in human language," *Proceedings of the National Academy of Sciences* **100**, 788–791(2003).

[98] H. S. Heaps, *Information Retrieval: Computational and Theoretical Aspects.* (Academic Press, New York NY, 1978).

[99] S. Bernhardsson, L. E. Correa da Rocha, P. Minnhagen, "The meta book and size-dependent properties of written language," *New Journal of Physics* **11**, 123015 (2009).

[100] J. B. Michel, *et al.*, "Quantitative Analysis of Culture Using Millions of Digitized Books," *Science* **331**, 176–182 (2011).

[101] *Google* n-gram project.

`http://ngrams.googlelabs.com`

[102] M. A. Nowak, *Evolutionary Dynamics: exploring the equations of life* (BelknapHarvard, Cambridge MA, 2006).

[103] M. A. Montemurro, P. A. Pury, "Long-range fractal correlations in literary corpora," *Fractals* **10**, 451–461 (2002).

[104] A. Corral, R. Ferrer i Cancho ,A. Diaz-Guilera "Universal complex structures in written language," e-print, arXiv:0901.2924v1 (2009).

[105] E. G. Altmann, J. B. Pierrehumbert, A. E. Motter, "Beyond word frequency: bursts, lulls, and scaling in the temporal distributions of words," *PLoS ONE* **4**, e7678 (2009).

[106] M. Sigman, G. A. Cecchi, "Global organization of the Wordnet lexicon," *Proceedings of the National Academy of Sciences* **99**, 1742–1747 (2002).

[107] M. Steyvers, M., J. B. Tenenbaum, "The large-scale structure of semantic networks: statistical analyses and a model of semantic growth," *Cognitive Science* **29** 41–78 (2005).

[108] E. Alvarez-Lacalle, B. Dorow, J.-P. Eckmann, E. Moses, "Hierarchical structures induce long-range dynamical correlations in written texts," *Proceedings of the National Academy of Sciences* **103**, 7956–7961 (2006).

[109] M. A. Nowak, N. L. Komarova, P. Niyogi, "Computational and evolutionary aspects of language," *Nature* **417**, 611–617 (2002).

[110] E. Lieberman, J.-B. Michel1, J. Jackson, T. Tang, and M. A. Nowak, "Quantifying the evolutionary dynamics of language," *Nature* **449**, 713–716 (2007).

[111] R. A. Blythe, "Neutral evolution: a null model for language dynamics," To appear in ACS *Advances in Complex Systems*.

[112] S. T. Piantadosi, H. Tily, and E. Gibson, "Word lengths are optimized for efficient communication," *Proceedings of the National Academy of Sciences* **108**, 3526–3529 (2011).

[113] A. M. Petersen, J. Tenenbaum,S. Havlin, and H. E. Stanley. In preparation.

[114] P. Klimek, W. Bayer, and S. Thurner, "The blogosphere as an excitable social medium: Richter's and Omori's Law in media coverage," *Physica A* **390**, 3870–3875 (2011).

[115] Y. Sano, K. Yamada, H. Watanabe, H. Takayasu, and M. Takayasu, "Empirical analysis of collective human behavior for extraordinary events in blogosphere," (preprint) arXiv:1107.4730 [physics.soc-ph].

[116] L. A. N. Amaral, S. V. Buldyrev, S. Havlin, H. Leschhorn, P. Maass, M. A. Salinger, H. E. Stanley, and M. H. R. Stanley, "Scaling Behavior in Economics: I. Empirical Results for Company Growth," *Journal de Physique I France* **7**, 621–633 (1997).

[117] D. Fu, F. Pammolli, S. V. Buldyrev, M. Riccaboni, K. Matia, K. Yamasaki, K., and H. E. Stanley, "The growth of business firms: Theoretical framework and empirical evidence," *Proceedings of the National Academy of Sciences* **102**, 18801–18806 (2005).

[118] M. H. R. Stanley, L. A. N Amaral, S. V. Buldyrev, S. Havlin, S., H. Leschhorn, P. Maass, M. A. Salinger, and H. E. Stanley, "Scaling behaviour in the growth of companies," *Nature* **379**, 804–806 (1996).

[119] D. Canning, L. A. N. Amaral, Y. Lee, M. Meyer, and H. E. Stanley, "Scaling the volatility of gdp growth rates," *Economic Letters* **60**, 335–341 (1998).

[120] L. A. N Amaral, S. V. Buldyrev, S. Havlin, M. Salinger, and H. E. Stanley, "Power Law Scaling for a System of Interacting Units with Complex Internal Structure," *Physical Review Letters* **80**, 1385–1388 (1998).

[121] M. Riccaboni, F. Pammolli, S. V. Buldyrev, L. Ponta, and H. E. Stanley, "The size variance relationship of business firm growth rates," *Proceedings of the National Academy of Sciences* **105**, 19595–19600 (2008).

[122] D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, and H. A. Makse, "Scaling laws of human interaction activity," *Proceedings of the National Academy of Sciences* **106**, 12640–12645 (2009).

[123] C. K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley and A. L. Goldberger, "Mosaic organization of DNA nucleotides," *Physical Review E* **49**, 1685 – 1689 (1994).

[124] K. Hu, Z. Chen, P. Ch. Ivanov, P. Carpena, and H. E. Stanley, "Effect of Trends on Detrended Fluctuation Analysis," *Physical Review E* **64**, 011114 (2001).

[125] E. G. Altmann, J. B. Pierrehumbert, A. E. Motter, "Niche as a determinant of word fate in online groups," *PLoS ONE* **6**, e19009 (2011).

[126] A. L. Barabási, "The origin of bursts and heavy tails in human dynamics," *Nature* **435**, 207–211 (2005).

[127] R. Crane, D. Sornette "Robust dynamic classes revealed by measuring the response function of a social system" *Proceedings of the National Academy of Sciences* **105**, 15649–15653 (2008) .

[128] D. Lazer, *et al.* "Computational social science," *Science* **323**: 721–723 (2009).

[129] V. Loreto, A. Baronchelli, A. Mukherjee, A. Puglisi, and F. Tria, "Statistical physics of language dynamics," *Journal of Statistical Mechanics* **2011**, P04006 (2011).

[130] A. Baronchelli, V. Loreto, and L. Steels, "In-depth analysis of the Naming Game dynamics: the homogenous mixing case," *International Journal of Modern Physics C* **19**, 785–812 (2008).

[131] A. Puglisi, A. Baronchelli, and V. Loreto, "Cultural route to the emergence of linguistic categories," *Proceedings of the National Academy of Sciences* **105**, 7936–7940 (2008).

[132] R. V. Solé, B. Corominas-Murtra, and J. Fortuny, "Diversity, competition, extinction: the ecophysics of language change," *Journal of the Royal Society Interface* **7**, 1647–1664 (2010).

[133] A. M. Petersen, F. Wang, S. Havlin, and H. E. Stanley, "Quantitative law describing market dynamics before and after interest-rate change," *Physical Review E* **81**, 066121 (2010).

[134] S. V. Buldyrev, M. Riccaboni, J. Growiec, H. E. Stanley, and F. Pammolli, "The growth of business firms: Facts and theory," *Journal of the European Economic Association* **5**, 574–584 (2007).

[135] B. Podobnik, D. Horvatic, A. M. Petersen, and H. E. Stanley, "Quantitative relations between risk, return, and firm size," *Europhysics Letters* **85**, 50003 (2009).

[136] Y. Liu, P. Gopikrishnan, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, "The Statistical Properties of the Volatility of Price Fluctuations," *Physical Review E* **60**, 1390–1400 (1999).

[137] Y. Lee, L. A. N. Amaral, D. Canning, M. Meyer, and H. E. Stanley, "Universal Features in the Growth Dynamics of Complex Organizations," *Physical Review Letters* **81**, 3275–3278 (1998).

[138] V. Plerou, L. A. N. Amaral, P. Gopikrishnan, M. Meyer, and H. E. Stanley, "Similarities between the growth dynamics of university research and of competitive economic activities" *Nature* **400**, 433–437 (1999).

[139] S. Picoli Jr., R. S. Mendes, L. C. Malacarne, E. K. Lenzi, "Scaling behavior in the dynamics of citations to scientific journals," *Europhysics Letters* **75**, 673–679 (2006).

[140] S. Picoli Jr., R. S. Mendes, "Universal features in the growth dynamics of religious activities," *Physical Review E* **77**, 036105 (2008).

[141] A. M. Petersen, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Persistence and Uncertainty in the Academic Career," *Proceedings of the National Academy of Sciences* (2011). DOI: 10.1073/pnas.1121429109 (2012).

[142] T. H. Keitt, H. E. Stanley "Dynamics of North American breeding bird populations" *Nature* **393**, 257–260 (1998).

[143] Redner, S. *A Guide to First-Passage Processes.* (Cambridge University Press, New York, 2001).

# Curriculum Vitae

# Curriculum Vitae

**Joel Tenenbaum**

## Office Address

590 Commonwealth Avenue
Department of Physics, Boston University
Boston, MA 02215, U.S.A.

**Office:** (617)353-8051
**Mobile:** (443)742-2148
**E-mail:** tenenbaum.joel@gmail.com

### EDUCATION

- 2011 (expected May 2012): Ph.D. (Physics), Boston University, Boston, MA 02215 USA
  (Thesis Advisor: H. Eugene Stanley).
- 2008: M. A. (Physics) Boston University, Boston, MA, 02215 USA
- 2006: B. A. (Physics and Mathematics with Music minor) Goucher College, Baltimore, MD 21204

### LIST OF PUBLICATIONS

1. Z. Zheng, K. Yamasaki, J. Tenenbaum, B. Podobnik, and H. E. Stanley. '"Scaling of Seismic Memory with Earthquake Size." Submitted September, 2011.
2. A. M. Petersen, J. Tenenbaum, S. Havlin, and H. E. Stanley. "Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death." **2**, March 15 2012, Scientific Reports.

3. J. Tenenbaum, S. Havlin, and H. E. Stanley. "Earthquake networks based on similar activity patterns." Submitted April, 2011, arXiv:1105.3415.

4. J. Tenenbaum, D. Horvati, S. C. Baji, B. Pehlivanovi, B. Podobnik, and H. E. Stanley. "Comparison between response dynamics in transition and developed economies." Physical Review E V **82**, October 8, 2010.

5. S. Dukan, J. Tenenbaum, J. Porembski, and K. Tata. "STM differential conductance of a disordered extreme type-II superconductor at high magnetic fields." Physical Review B V **82**, October 4, 2010.

6. B. Podobnik, D. Horvatic, J. Tenenbaum, and H. E. Stanley. "Asymmetry in power-law magnitude correlations." Physical Review E: Rapid Communications V **80**, July 17, 2009.

7. J. Tenenbaum and S. Dukan. "Differential Conductance of Extreme Type-II Superconductors in High Magnetic Fields." Proceedings of the National Conference of Undergraduate Research (NCUR), Lexington, VA. April, 2005

**PRESENTATIONS**

- "Differential Conductance of Extreme Type-II Superconductors in High Magnetic Fields"
  - *National Conference of Undergraduate Research (NCUR)*, Lexington, VA. April, 2005.
  - *Posters on the Hill, Council on Undergraduate Research.* Washington, D.C., April, 2005

- "Correlation Networks of Earthquakes" *APS March Meeting 2009*, American Physical Society. Pittsburgh, PA, March 19, 2009.

- "Correlation Networks of Earthquakes" *International Workshop on Network Science 2009.* Venice, July 1, 2009.

- "Can Networks Help Understand Earthquake Physics?" *International Conference on Statistical Physics of the International Union for Pure and Applied Physics (IUPAP) 2010.* Cairns, Australia, July 29, 2010.

- "The Sign Effect in Emerging Markets: The Inherent Instability of Bad News". *HES70: Horizons in Emergence & Scaling.* Poster session. Boston, March, 2011.

- "The Sign Effect in Emerging Markets: The Inherent Instability of Bad News". *APS March Meeting 2011*, American Physical Society. Dallas, TX, March 21, 2011.

- "The Growth Dynamics of Words: How Historical Context Shapes the Competitive Linguistic Environment". *APS March Meeting 2012*, American Physical Society. Boston, Feb 28, 2012.

## TEACHING EXPERIENCE

**Teaching Fellow, Boston University, Boston, MA** 2006-Present

- Lead discussion sections and labs, coordinating with undergraduate courses
- Explained physical and mathematical concepts
- Demonstrating solving problems in real-time on blackboard

**Tutor and Supplemental Instructor, Goucher College, Towson, MD** 2002-2006

- Taught supplemental class sessions for first year and second year physics
- Tutored through Academic Center for Excellence in both physics and math

**Mentor, BUILD Greater Boston, Boston, MA** 2011-2012

- Taught entrepreneurial skills to high school students in a mentorship capacity

**Tutor and Substitute Teacher, Boston University Academy, Boston, MA** 2007-Present

- Taught problem solving and critical thinking to high school physics, mathematics, and chemistry students

**WRITING EXPERIENCE**

"How it feels to be sued for $4.5m", *The Guardian*, 27 July 2009.

"Filesharer Joel Tenenbaum's trial diary: Part one", *The Guardian*, 6 August 2009.

"Filesharer Joel Tenenbaum's trial diary: Part two", *The Guardian*, 7 August 2009.

"Filesharer Joel Tenenbaum's trial diary: Part three", *The Guardian*, 10 August 2009.

"Filesharer Joel Tenenbaum's trial diary: Part four", *The Guardian*, 11 August 2009.

"Joel Tenenbaum: a year on from being sued for $4.5m", *The Guardian*, 9 November 2010.