Generalized entropies and the similarity of texts

J. Stat. Mech. (2017) 014002

(http://iopscience.iop.org/1742-5468/2017/1/014002)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

Special Issue on Statphys 26

Generalized entropies and the similarity of texts

Eduardo G Altmann[1,2], **Laércio Dias**[1] **and Martin Gerlach**[1,3]

[1] Max Planck Institute for the Physics of Complex Systems, D-01187 Dresden, Germany
[2] School of Mathematics and Statistics, University of Sydney, 2006 NSW, Australia
[3] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA
E-mail: eduardo.altmann@sydney.edu.au

Received 14 November 2016
Accepted for publication 11 December 2016
Published 27 January 2017

Online at stacks.iop.org/JSTAT/2017/014002
doi:10.1088/1742-5468/aa53f5

Abstract. We show how generalized Gibbs–Shannon entropies can provide new insights on the statistical properties of texts. The universal distribution of word frequencies (Zipf's law) implies that the generalized entropies, computed at the word level, are dominated by words in a specific range of frequencies. Here we show that this is the case not only for the generalized entropies but also for the generalized (Jensen–Shannon) divergences, used to compute the similarity between different texts. This finding allows us to identify the contribution of specific words (and word frequencies) for the different generalized entropies and also to estimate the size of the databases needed to obtain a reliable estimation of the divergences. We test our results in large databases of books (from the google n-gram database) and scientific papers (indexed by Web of Science).

Keywords: scaling in socio-economic systems

J. Stat. Mech. (2017) 014002

1742-5468/17/014002+13$33.00

Contents

1. Introduction

Generalized entropies, such as the Renyi and Tsallis entropies, have been studied in different aspects of statistical physics [1, 2] and non-linear dynamics [3]. In information theory, these entropies are viewed as a generalizations of the Shannon entropy that are potentially useful in particular problems. Many problems require the comparison of the divergence (or, its opposite, the similarity) between two or more signals, a problem that can be quantified through the use of divergence measures based on generalized (joint) entropies, e.g. in analysis of DNA sequences [4] or image processing [5].

A traditional and increasingly important application of information theory is the analysis of (signals based on) natural language [6–11]. This analysis often happens at the level of words, i.e. in which each word (type) is considered a different symbol of analysis. One important statistical feature in the statistical analysis of word frequencies is the existence of linguistic laws [12], i.e. statistical regularities observed in a variety of databases. The most famous case is Zipf's law, which specifies how the frequencies of words are distributed [13–16].

In this paper we explore the implications of linguistic laws to the computation of information-theoretic measures in written text. While information-theoretic approaches typically measure the similarity of an ensemble of words (the vocabulary), we show how

J. Stat. Mech. (2017) 014002

generalized entropies can be used to assess the influence of individual words to these (global) measures, providing a bridge to the studies on evolution of language following trajectories of individual words [17, 18]. In particular, we show how the contribution of individual words, appearing in different scales of frequency, vary in the different generalized entropies. We explore the implications of our findings to two problems: (i) the best generalized entropy for highlighting the contribution of physics keywords; and (ii) determining how large a given database has to be in order obtain sufficient coverage/sampling of the generalized entropies.

2. Basic concepts

We are interested in extracting information about written documents based on the number of times N_i each word $i = 1, \ldots, M$ appears in each database. For each database, we denote by f_i the frequency of the word i (i.e. $f_i \equiv N_i / \sum_{i=1}^{M} N_i$), which we consider to be an estimator of the probability p_i of occurrence of this word in the generative process underlying the production of the texts. We say that the word i has rank r if it is the rth most frequent word.

2.1. Zipf's law

Different databases show similar distributions of word frequencies, a statistical regularity also known as Zipf's law. While Zipf originally proposed the simple relationship $f(r) \propto 1/r$, more recent analysis in large text databases suggest that the data is better described by a double power-law (dp) distribution [14, 19–21]

$$f(r) = F^{(\mathrm{dp})}(r; \gamma, b) = C \begin{cases} r^{-1}, & r < b \\ b^{\gamma-1} r^{-\gamma} & r \geqslant b, \end{cases} \tag{1}$$

where b and γ are free parameters, $C = C(\gamma, b)$ is the normalization constant (which can be approximated as $C \approx 1/(G^1_{b-1} + 1/(\gamma - 1))$, and $G^a_b \equiv \sum_{r=1}^{b} r^{-a}$ is the bth generalized Harmonic number [22]. The more common single-power-law distribution is recovered for $b \to 1$ and our results below apply in this limit as well. In plots and numerical calculations we use the distribution (1) with $b = 7873, \gamma = 1.77$, and $C = 0.0922$, values obtained in [14] for english books published in different centuries. In figure 1 we show that the modified Zipf's law indeed provides good account of different databases.

2.2. Generalized entropies

In line with the long-tradition of information theory, we use entropies to quantify the amount of information contained in written texts. Here we consider the generalized entropy of order α [23]

$$H_\alpha(\boldsymbol{f}) = \frac{1}{1-\alpha} \left(\sum_{i=1}^{M} (f_i)^\alpha - 1 \right), \tag{2}$$

J. Stat. Mech. (2017) 014002

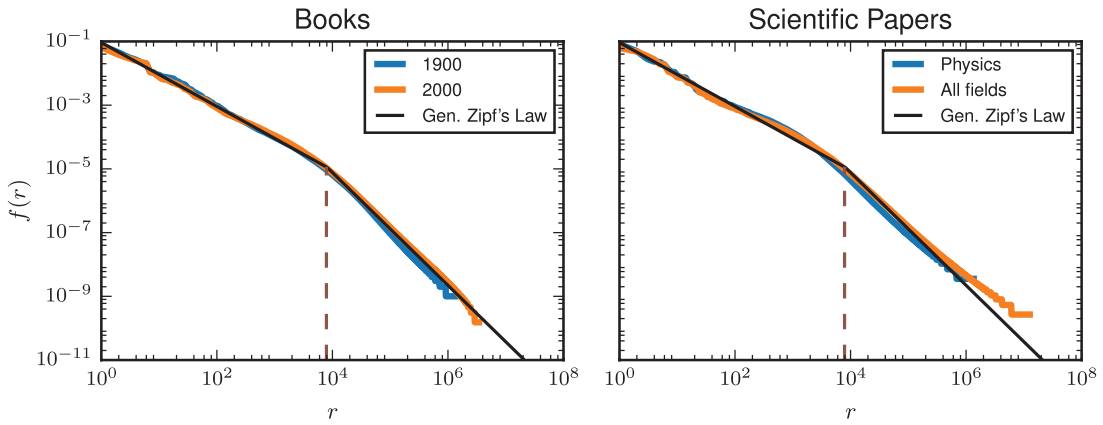

Figure 1. Frequencies of words are distributed over a variety of scales and are well described by the modified Zipf's law. The (thin) black line corresponds to equation (1) with $b = 7873$ and $\gamma = 1.77$ [14]. The (thick) colored lines correspond to the frequency of words obtained in different databases. (Left) Results for books published in the years 1900 and 2000 (see legend), as provided by the google n-gram database; (right) results for the abstract of scientific papers indexed in the Web of Science between 1991 to 2014 (in Physics and in all fields, see legend).

where $\boldsymbol{f} = (f_1, f_2, \ldots, f_M)$, the sum runs over all words for which $f_i \neq 0$, and α is a free parameter yielding a spectrum of entropies. For $\alpha = 1$ we recover the Gibbs–Shannon entropy, i.e. $H_{\alpha=1} = -\sum_i f_i \log f_i$. In Physics, equation (2) is known as Tsallis entropy [1, 2] and has been proposed as a (non-extensive) generalization of the traditional statistical mechanics.

2.3. Divergence measures

We are particularly interested in using H_α to quantify the distance (or dissimilarity) between different databases. Here we focus on the generalized Jensen–Shannon divergence [24]

$$D_\alpha(\boldsymbol{p}, \boldsymbol{q}) = H_\alpha\left(\frac{\boldsymbol{p}+\boldsymbol{q}}{2}\right) - \frac{1}{2}H_\alpha(\boldsymbol{p}) - \frac{1}{2}H_\alpha(\boldsymbol{q}), \tag{3}$$

where $\boldsymbol{p}$ and $\boldsymbol{q}$ are the word frequencies of the two databases and $\boldsymbol{p}+\boldsymbol{q} = \sum_i p_i + q_i$ is obtained summing over all symbols for which either $p_i \neq 0$ or $q_i \neq 0$. We focus on D_α because $\sqrt{D_\alpha}$ can be shown to be a metric for $0 < \alpha \leqslant 2$, i.e. it is positive $D_\alpha \geqslant 0$ (with $D_\alpha = 0$ if and only if $\boldsymbol{p} = \boldsymbol{q}$), symmetric $D_\alpha(\boldsymbol{p}, \boldsymbol{q}) = D_\alpha(\boldsymbol{q}, \boldsymbol{p})$, and $\sqrt{D_\alpha}$ satisfies the triangular inequality [4, 25, 26]. We expect our main results to apply also to other quantities obtained from $H_\alpha(\boldsymbol{p}, \boldsymbol{q})$, $H_\alpha(\boldsymbol{p})$, and $H_\alpha(\boldsymbol{q})$, such as the generalized mutual information and Kullback–Leibler divergence [27]. The usual ($\alpha = 1$, Jensen–Shannon) divergence is a traditional method in different statistical analysis of natural language [6]. For generalized entropies, increasing (decreasing) α one increases (decreases) the weight of the most frequent words allowing for different insights into the relationship between the databases [28].

J. Stat. Mech. (2017) 014002

3. Effect of Zipf's law on generalized measures

The goal of this paper is to investigate the consequences of known properties of word statistics to the computation of generalized entropic measures. For instance, the number of different words is virtually unbounded and therefore we should carefully consider finite-size effects and the role played by the number of observed symbols in our analysis [9, 28]. More specifically, we explore the consequences of Zipf's law—as reviewed in section 2.1—to the computation of the information-theoretic measures based on H_α —reviewed in sections 2.2 and 2.3. In [28] we have shown that Zipf's law implies that finite-size estimators of H_α and D_α scale very slowly with database size. Here we focus on the contribution of individual words to H_α and D_α, showing how different frequency ranges dominate the estimation for different values of α.

3.1. Entropy H_α

The entropy (2) is uniquely defined by the frequency of the words $\boldsymbol{f}$. From the double power-law (dp) frequency distribution, equation (1), we obtain

$$H_\alpha^{(\mathrm{dp})} \equiv \frac{1}{1-\alpha}\left(\sum_{r=1}^{\infty}(F_{\mathrm{dp}}(r))^\alpha - 1\right) = \frac{1}{1-\alpha}(C^\alpha(h_1+h_2)-1), \tag{4}$$

with

$$h_1 = \sum_{r=1}^{b-1} r^{-\alpha} \equiv G_{b-1}^{\alpha} \text{ (generalized Harmonic number)},$$

and

$$h_2 = b^{\alpha(\gamma-1)}\sum_{r=b}^{\infty} r^{-\alpha\gamma} = b^{\alpha(\gamma-1)}(\zeta(\alpha\gamma) - G_{b-1}^{\alpha\gamma}) \approx \frac{b^{1-\alpha}}{\alpha\gamma-1},$$

where $\zeta(a)$ is the Riemann zeta function and the right hand side is obtained approximating the sum by the integral and is valid for $\alpha > 1/\gamma$ (where $H_\alpha < \infty$). The divergence of H_α for $\alpha \leqslant 1/\gamma$ appears because the sum/integral diverges for $r \to \infty$ (i.e. for a growing number of different words). A comparison between H_α in real data and $H_\alpha^{(\mathrm{dp})}$ is shown in figure 2(a). The difference between the theory and the data for $\alpha \lessapprox \alpha_c = 1/\gamma$ is due to the finite number of symbols in the database. This is a finite-size effect that depends sensitively on the size of the database used to estimate $\boldsymbol{f}$.

We now focus on the contribution of individual words for H_α. To do that, we take advantage of the fact that H_α can be written as a sum over different words and consider the ratio

$$R(r) = \frac{\sum_{r'=1}^{r}(f_{r'})^\alpha}{\sum_{r'=1}^{\infty}(f_{r'})^\alpha} \tag{5}$$

as a proxy for the contribution of the first r terms to the computation of H_α. For the case of the double power-law distribution $f_r = F_{\mathrm{dp}}(r)$, we obtain that

J. Stat. Mech. (2017) 014002

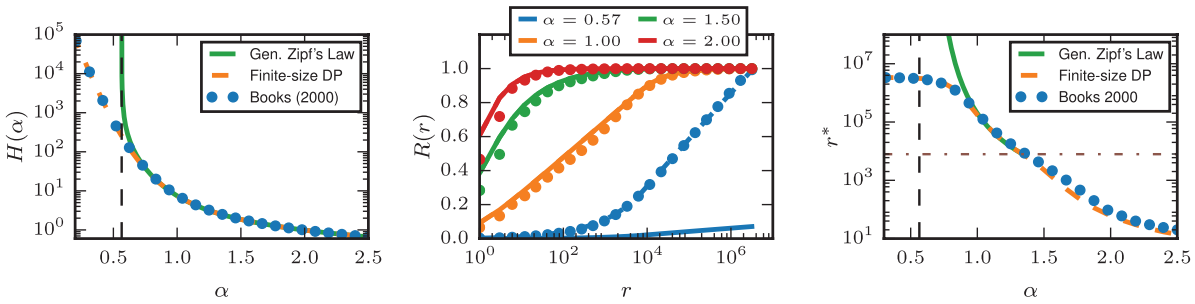

Figure 2. Contribution of the r most frequent words to the estimation of the generalized entropy H_α. Symbols are the results obtained for the data (books published in the year 2000). Lines are the theoretical predictions from the double-power-law distribution (1) with the same number of words as in the data (dashed line, finite-size DP) and with infinite support (solid line, obtained analytically). (a) H_α as a function of α, solid line corresponds to equation (4); (b) contribution of the r most frequent words measured by the ratio $R_\alpha(r)$ given in equation (5), solid lines correspond to equation (6); and (c) the rank r^* for which $R_\alpha(r = r^*) = 99\%$, solid line corresponds to equations (8)–(9).

$$(h_1 + h_2)R^{(\mathrm{dp})}(r) = \begin{cases} \displaystyle\sum_{r'=1}^{r} r'^{\,-\alpha}, & = G_r^\alpha, \text{ for } r < b \\ \displaystyle\sum_{r'=1}^{b-1} r'^{\,-\alpha} + b^{\alpha(\gamma-1)} \sum_{r'=b}^{r} r'^{\,-\gamma\alpha} & = G_{b-1}^\alpha + b^{\alpha(\gamma-1)}(G_r^{\alpha\gamma} - G_{b-1}^{\alpha\gamma}) \text{ for } r \geqslant b. \end{cases} \tag{6}$$

For $r > b$ we can approximate the sum $\sum_{r'=b}^{r} r'^{\,-\gamma\alpha}$ by an integral and obtain

$$R^{(\mathrm{dp})}(r) \approx \frac{1}{h_1 + h_2}\left(h_1 + \frac{b^{\alpha(\gamma-1)}}{\alpha\gamma - 1}(b^{1-\alpha\gamma} - r^{1-\alpha\gamma})\right). \tag{7}$$

In figure 2(b) we show the dependence of R and $R^{(\mathrm{dp})}$ on r for different values of α. A deviation due to finite-size effects is again observed when $\alpha \to 1/\gamma$ (finite database size).

The analysis of R reveals a convergence that varies dramatically with α (see also [9, 28]), suggesting that for different α's different ranges in f contribute to H_α. One quantity of interest is the rank r_q^* so that $r \leqslant r_q^*$ accounts for a fraction q of the effect, e.g. for $q = 0.99$ we have that $R(r_q^*) = 0.99$ meaning that the first r_q^* terms are responsible for 99% of the total $\sum_r f_i^\alpha$. For small q or large α, such that $r_q^* < b$, r_q^* is obtained from the first line of equation (6) as the solution of

$$G^\alpha_{r_q^*} = q. \tag{8}$$

For large q or small α, such that $r_q^* > b$, r_q^* can be obtained explicitly from equation (7) as

$$r_q^*(\alpha) = \left(b^{1-\alpha\gamma} - \frac{\alpha\gamma - 1}{b^{\alpha(\gamma-1)}}(qh_2 - (1-q)h_1)\right)^{1/(1-\alpha\gamma)}. \tag{9}$$

The estimations (8) and (9), which are based on the double power-law distribution (1), and the results obtained in the data are shown in figure 2(c). We see that for $\alpha = 1$

J. Stat. Mech. (2017) 014002

one typically needs around 200 000 different word types in order to obtain 99% of the asymptotic value of R. This number quickly decays with α so that for $\alpha = 2$, the 100 most frequent words lead to the same relative contribution and therefore all other words are irrelevant in practice.

3.2. Divergence D_α

The divergence D_α defined in equation (3) quantifies how dissimilar two databases are ($\boldsymbol{p}$ and $\boldsymbol{q}$) and the distribution of frequencies in these databases alone does not specify D_α. Still, we expect the general shape of Zipf's law in equation (1) to affect the statistical properties of D_α. Here we explore this connection by following steps similar to those performed in the previous section for H_α. To do this, it is convenient to introduce the relative coordinates f_i, Δ_i, where $f_i = (p_i + q_i)/2$ and $\Delta_i = |p_i - q_i|/2$, such that:

$$D_\alpha(\boldsymbol{p}, \boldsymbol{q}) = D_\alpha(\boldsymbol{f}, \boldsymbol{\Delta}) = \sum_i \frac{1}{1-\alpha}\left((f_i)^\alpha - \frac{1}{2}(f_i + \Delta_i)^\alpha - \frac{1}{2}(f_i - \Delta_i)^\alpha\right) \equiv \sum_r D_\alpha(r). \tag{10}$$

This equation emphasizes that D_α is computed as a sum over a contribution $D_\alpha(r)$ of different words ranked by r. We order the words according to the rank r of the word in $\boldsymbol{f}$, i.e. if a word has rank r' it means that there are exactly $r' - 1$ other words for which the average frequency $f_r = (p_r + q_r)/2 > f_{r'} = (p_{r'} + q_{r'})/2$.

The relative contribution $\mathbb{R}(r)$ of the top r words to D_α is given by

$$\mathbb{R}(r) = \frac{\sum_{r'=1}^{r} D_\alpha(r')}{D_\alpha} = \frac{\sum_{r'=1}^{r}\left((f_{r'})^\alpha - \frac{1}{2}(f_{r'} + \Delta_{r'})^\alpha - \frac{1}{2}(f_{r'} - \Delta_{r'})^\alpha\right)}{\sum_{r'=1}^{\infty}\left((f_{r'})^\alpha - \frac{1}{2}(f_{r'} + \Delta_{r'})^\alpha - \frac{1}{2}(f_{r'} - \Delta_{r'})^\alpha\right)}, \tag{11}$$

which is analogous to equation (5) but in this case $D_\alpha(r)$ is not necessarily monotonically decaying with r. We finally define $\mathbb{r}_q^*$ as the rank at which a fraction q of the total D_α is achieved, i.e. $\mathbb{R}(\mathbb{r}_q^*) = q$.

Figure 3 shows our analysis of the divergence (D_α, $\mathbb{R}(r)$, and $\mathbb{r}_q^*$) for two pairs of databases (Books2000–Books1900 and Books2000–Physics, see caption of figure 1 for details on the data). The left panel shows that the divergence D_α for Books2000–Physics is systematically larger than for Books2000–Books1900 suggesting that stylistic and topical differences between books and scientific papers are more significant than historical changes in the language throughout the 20th century. The most striking feature of figure 3 is the similarity between the results obtained with different data (e.g. the variation across the databases is much smaller than the variation across α or r). Furthermore, the general behavior observed for D_α resembles the results shown in figure 2 for H_α, which were analytically computed from the word-frequency distribution (1). The D_α-observation, however, depends not only on the word frequencies f_i but also on the variation Δ_i across databases. Next we consider two very simplistic models for Δ_i in order to understand these observations.

3.2.1. Constant relative fluctuation. A simple assumption is that the relative fluctuations across databases are the same for each word independent of its frequency, in which case $\boldsymbol{\Delta}$ is proportional to the average frequencies $\boldsymbol{f}$ and thus

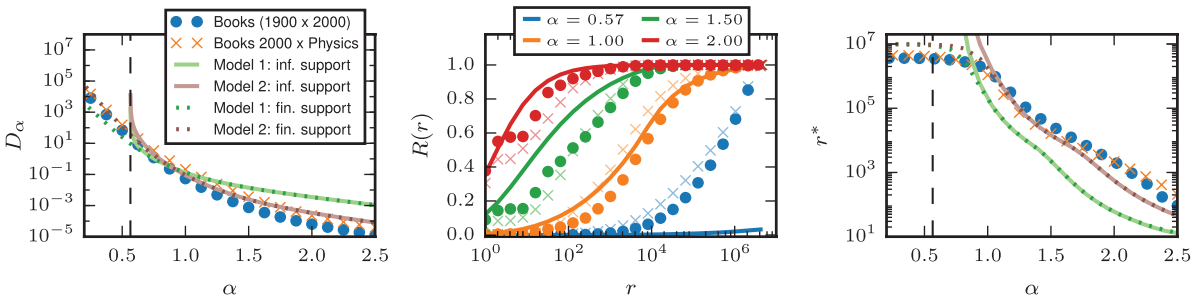

Figure 3. Contribution of the r most frequent words to the estimation of the generalized divergence D_α. Symbols are the results obtained for the data: books published in 1900 versus books published in 2000 (dots) and books published in 2000 versus abstracts of Web of Science papers (crosses). Lines are the theoretical predictions from the double-power-law distribution (1) with infinite support assuming $\Delta_i \propto f_i$, equation (12) (light solid line, model 1), and $\Delta_i \propto f_i \log f_i$, equation (16) (dark solid line, model 2). (a) D_α as a function of α; (b) contribution of the r-most frequent words (ranked by the average frequency); and (c) the rank r^* for which $R_\alpha(r = r^*) = 99\%$.

$$\frac{\Delta_i}{f_i} = A. \tag{12}$$

In this case we obtain from (10) that

$$D_\alpha = \left(1 - \frac{1}{2}(1-A)^\alpha - \frac{1}{2}(1+A)^\alpha\right)\frac{1}{1-\alpha}\sum_r (f_r)^\alpha \tag{13}$$

$$= \left(1 - \frac{1}{2}(1-A)^\alpha - \frac{1}{2}(1+A)^\alpha\right)\left(H_\alpha(\boldsymbol{f}) + \frac{1}{1-\alpha}\right) \tag{14}$$

$$\approx \frac{\alpha(1-\alpha)}{2}A^2\left(H_\alpha(\boldsymbol{f}) + \frac{1}{1-\alpha}\right), \tag{15}$$

where the approximation is valid for $A \ll 1$. Now we notice that $\boldsymbol{f}$ is the word frequency distribution of the combined database and that therefore it should also be well approximated by the generalized Zipf's law (1). Even if this model is too simplistic to account for the observed D_α (see dotted line in the left panel of figure 3), it shows how the statistical properties of D_α and of H_α can be connected to each other.

3.2.2. Log-corrected fluctuations. In order to get some insights on the reason for the failure of the previous model, we look at the empirical relative fluctuation $\frac{\Delta_i}{f_i}$ for the two pairs of databases described above. The results in figure 4 show two features: an expected large fluctuation around different words and a surprising decay of relative fluctuation with f_i. The roughly linear decay in the semi-logarithmic plot suggests that an improvement of equation (12) is obtained including a logarithmic correction

J. Stat. Mech. (2017) 014002

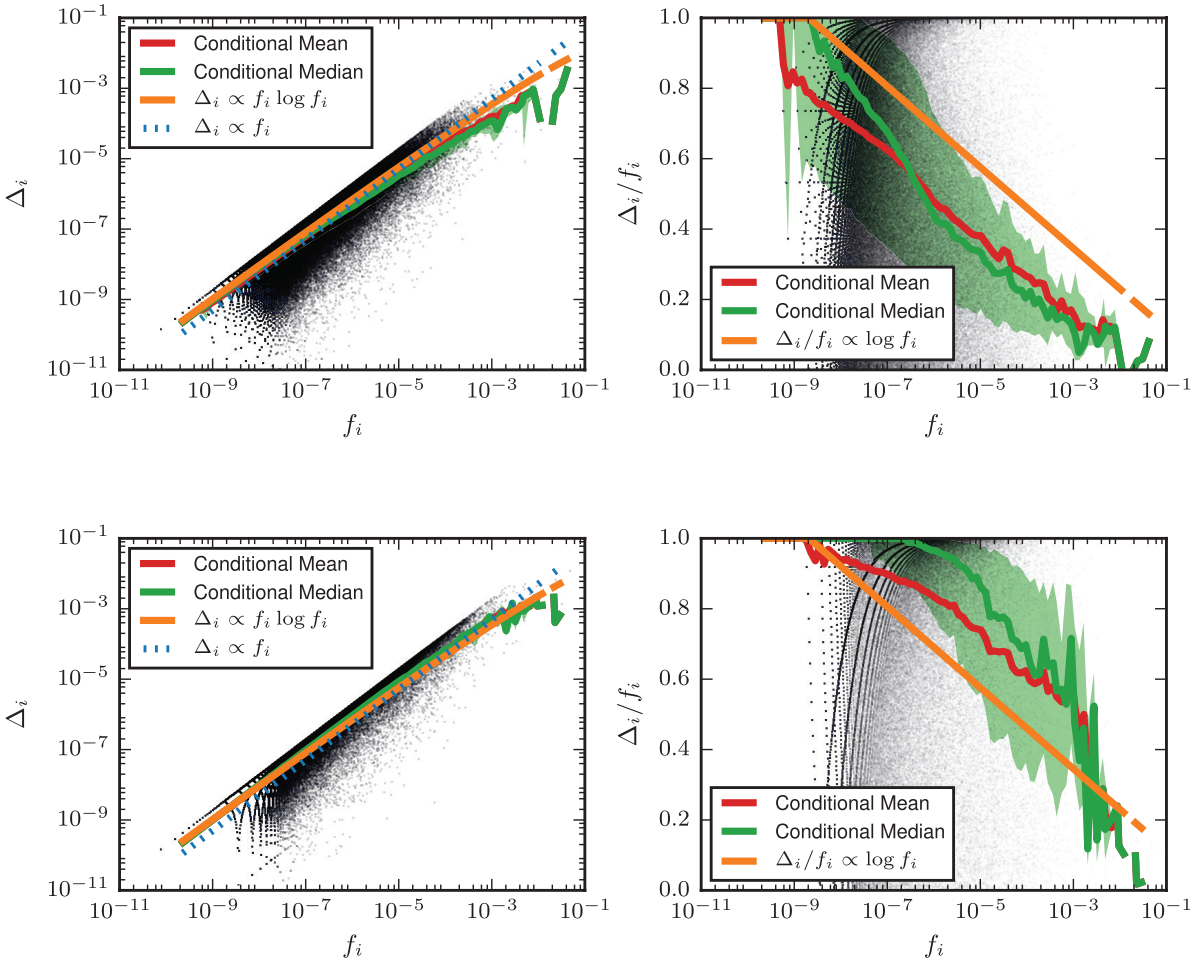

Figure 4. Relation between relative $\Delta_i = |p_i - q_i|/2$and average $f_i = (p_i + q_i)/2$ frequency. Mean and median (conditioned on window in f_i) are shown for divergences between books published in the year 1900 and 2000 (top panels) and books published in 2000 and abstracts from WoS (bottom panels). Shaded region correspond to 25- and 75-percentile. Approximations for the conditional mean are given by $\Delta_i/f_i = 0.5$ (dotted line) and $\Delta_i/f_i = -0.05 \log f_i$ (dashed line).

as $\Delta_i/f_i \propto \log f_i$. Since Δ_i is bounded from above by fi (i.e. $\Delta_i \leqslant f_i$) we introduce a lower cutoff frequency in our log-corrected model

$$\frac{\Delta_i}{f_i} = \begin{cases} a \log f_i/f_{\max} & , f > f_{\max} e^{1/a} \\ 1 & , f \leqslant f_{\max} e^{1/a} \end{cases}, \tag{16}$$

where we empirically find that $f_{\max} = 1$ and $a = -0.05$ capture the main qualitative behaviour shown in figure 4.

The log-corrected model, obtained combining equation (16) with the generalized Zipf's law (1), provides a much better account of the results in the three panels of figure 3. This shows that the weak dependence of the relative fluctuations on the frequency is crucial in order to understand the results in figure 3.

doi:10.1088/1742-5468/aa53f5

J. Stat. Mech. (2017) 014002

4. Implication of our results

4.1. Keywords in physics

Our results shows that the Zipf's law is responsible for the general statistical properties of both H_α and D_α. One consequence of this result is that the contribution of (a set of) particular words is also pre-determined by Zipf's law and depends largely on the range of frequencies of the words. Consider the problem of comparing the divergence between the corpus of scientific papers in Physics to a general corpus of books written in English. One of the effects one may want to capture when computing D_α is the over-representation of Physics-related words in the database of Physics articles, i.e. the fact that $p_i > q_i$ for words i related to Physics. We denote this set of words as Physics keywords. This is not the only effect contributing to the divergence D_α between the texts, e.g. stylistic effects affecting the most frequent words (so-called stopwords) may also be relevant. Here we wish to quantify the effect of Physics keywords to D_α in comparison to a set of stopwords.

The key insight that connects this problem to our results is that Physics keywords are typically distributed in a specific range of frequencies. For instance, we compiled a list of 318 Physics keywords from all words appearing in the PACS system (removing a list of common stop words). As illustrated in the figure 5(left panel) the words range from *electron*—with rank $r_i \approx 100$ and frequency of one every thousand words $f_i \approx 10^{-3}$—to *gravitation*—with rank $r_i \approx 2000$ and frequency of one every hundred thousand words $f_i \approx 10^{-5}$. Most Physics keywords lie in between these two frequencies. By increasing α from $\alpha = \alpha_c = 1/\gamma \approx 0.56$ one moves from a configuration in which D_α and H_α are dominated by the least frequent words to a configuration in which D_α and H_α are determined mostly by the most frequent stopwords (e.g. for $\alpha > 2$). Indeed, the results in figure 5(right panel) confirm that the contribution of the Physics keywords has a maximum around $\alpha \approx 1.4$. At the maximum, these 318 keywords contribute with more than 10% of the total value of D_α. This value is comparable to the contribution of the 10 most frequent words (stopwords) at the same value of α. The contribution of the stopwords quickly increases with α and completely dominates D_α for $\alpha \gtrapprox 2.0$.

4.2. How large does my database have to be?

When computing H_α and D_α one usually aims at characterizing the properties of the source (stochastic process) underlying the data. Stationarity and ergodicity of this process imply that computed values should converge for increasing database size. In practice, we are not interested in results which depend mainly on the size of the database, and that change dramatically with the amount of available data. Below we show how our results allow for an estimation of the database size required to provide a reliable estimation of D_α.

The most important effect of changing the database size is to increase the number of different words found in the databases. This simple observation, the cornerstone of our analysis, has two ramifications. First, it implies that a necessary condition for a robust estimation of D_α is that $M > \mathbb{r}^*_{q \longrightarrow 1\_}$, i.e. the number of observed different words M needs to be larger than the number of ranks r needed to estimate a fraction $q \lessapprox 1$ of D_α. Second, a connection to the size of the database N (measured in number of word tokens) is possible through Heaps' law, which states that the number of different words grows sublinear with the total number of words, $M \sim N^{1/\gamma}$ [29, 30]. In figure 6 we present the result of

J. Stat. Mech. (2017) 014002

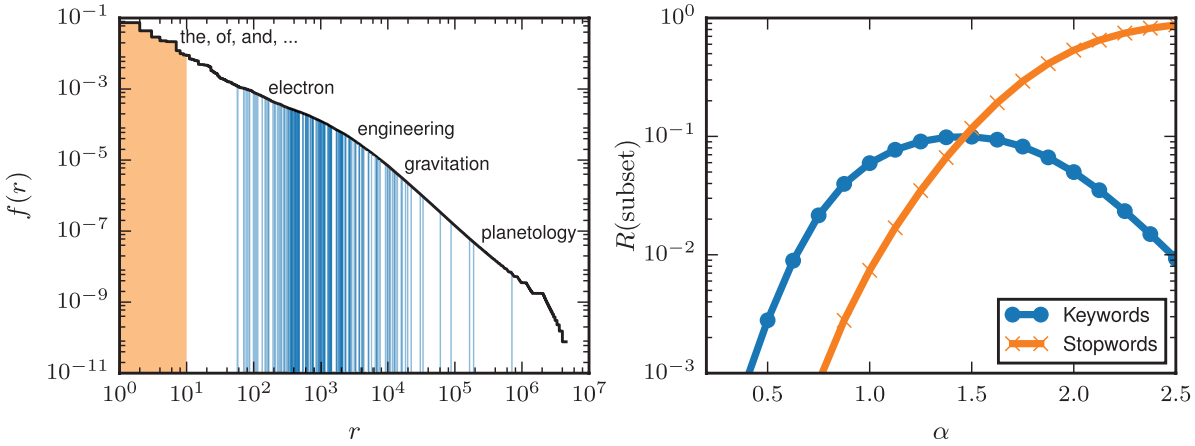

Figure 5. Contribution of subsets of words to the divergence D_α. Results are shown for a list of 318 physics keywords (see text) and a list of the 10 most frequent stopwords (the, of, and, in, to, a, is, for, that, with). (Left) Position of keywords and stopwords in the rank-frequency distribution. (Right) Fraction of the generalized divergence D_α from words belonging to the list of keywords and the list of stopwords as a function of α.

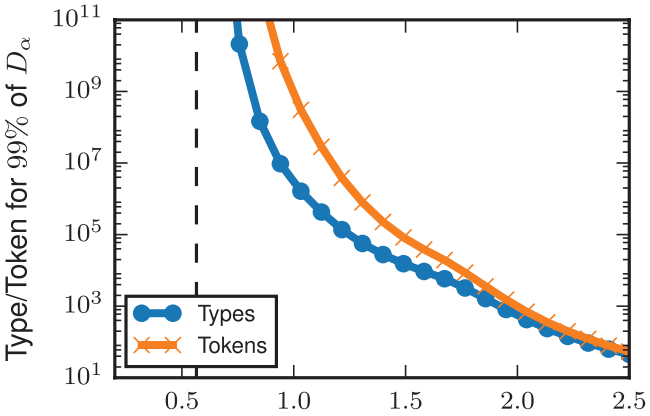

Figure 6. Database size necessary to observe 99% of D_α. The curve for the number of different words (types) M was computed from $\mathbb{r}$ as in figure 3. The relationship $M \sim N^{1/\gamma}$ to the size of the database N (number of tokens) was obtained from a Poisson null model assuming a double power-law Zipfian distribution, as in [31]. For comparison, the typical book size in Project Gutenberg is $N \approx 10^5$, implying that D_α between two books can typically be computed only for $\alpha > 1.5$.

this analysis, in which $\mathbb{r}^*_{q=0.99}$ was obtained from the double-power-law distribution with log-corrected fluctuations (as in figure 3) and the Heaps' law relationship derived in [31].

5. Discussion and conclusions

The main message of this paper is that the characteristic shape of word-frequency distributions (f_r following Zipf's law) plays a dominant role in the properties of information-theoretic measures computed in texts at the level of words. While there is a one-to-one

J. Stat. Mech. (2017) 014002

relationship between f_r and entropies H_α—given in equation (4)—here we showed that a close connection exists also between f_r and measures intended to compare databases such as D_α, a result that presumably extends also to other measures such as the mutual information and Kullback–Leibler divergence. The influence of f_r occurs not only in the convergence of finite-size estimators, as reported previously in [9, 28], it affects the value of D_α and the weight of the contributions of words in different frequency ranges. This connection relies not only on the universality of f_r but also on our empirical finding that, for different pairs of databases, the relative fluctuations decay with the logarithm of the frequency, see equation (16) and figure 4.

The finding that Zipf's law directly controls the expected weights of contribution of different words provides a further motivation for our choice of using generalized entropies H_α. The variation of the free parameter α effectively tunes the range of frequency of the words that contribute to H_α and D_α: for large α (e.g. $\alpha = 2$) only the most frequent words contribute, while for $\alpha < 1$ the results are dominated by the least frequent words. From an example based on 318 keywords in Physics, we obtain that these words contribute with 6% of $D_{\alpha=1}$, 10% of $D_{\alpha=1.4}$, but only 5% of $D_{\alpha=2}$. Words in different frequency ranges have different semantic and syntactic properties so that the variation of α can characterize also different types of divergencies between the databases.

As α is reduced and approaches (from above) the critical value $\alpha = 1/\gamma$, where γ is the exponent of Zipf's law defined in equation (1), the convergence of H_α and D_α becomes extremely slow and increasingly large text sizes are needed for a robust estimation (see figure 6). For instance, for the usual Jensen–Shannon divergence $D_{\alpha=1}$ we estimate that databases of size $\approx 10^8$ tokens ($\approx$200 books or $\approx 10^6$ word types) is needed while for $\alpha = 0.6$ the size grows dramatically to the unrealistic number of $\approx 10^{20}$ tokens ($\approx 2\,10^{14}$ books or $\approx 10^{16}$ word types). For $\alpha < 1/\gamma \approx 0.56$ there is no convergence and therefore these quantities are not properly defined. This is one of the most dramatic consequences of Zipf's law and reflects the effectively unbounded number of different symbols (vocabulary) in which H_α is computed.

Acknowledgments

L D received financial support from CNPq/Brazil through the program 'Science without Borders'. E G A and M G are grateful to F Font-Clos for helpful discussions on the subject of this manuscript. We thank Margit Palzenberger and the Max Planck Digital Library for providing access to the Web of Science dataset used in this paper.

References

[1] Gell-Mann M and Tsallis C 2004 *Nonextensive Entropy: Interdisciplinary Applications* (Oxford: Oxford University Press)
[2] Tsallis C 2009 *Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World* (New York: Springer)
[3] Kantz H and Schreiber T 2003 *Nonlinear Time Series Analysis* (Cambridge: Cambridge University Press)
[4] Grosse I, Bernaola-Galván P, Carpena P, Román-Roldán R, Oliver J and Stanley H 2002 *Phys. Rev.* E **65** 041905
[5] He Y, Hamza A B and Krim H 2003 *IEEE Trans. Signal Process.* **51** 1211–20
[6] Manning C and Schütze H 1999 *Foundations of Statistical Natural Language Processing* (Cambridge, MA: MIT Press)

J. Stat. Mech. (2017) 014002

[7] Boyack K W, Newman D, Duhon R J, Klavans R, Patek M, Biberstine J R, Schijvenaars B, Skupin A, Ma N and Börner K 2011 *PLoS One* **6** e18029
[8] Masucci A P, Kalampokis A, Eguíluz V M and Hernández-García E 2011 *PLoS One* **6** e17333
[9] Dodds P S, Harris K D, Kloumann I M, Bliss C A and Danforth C M 2011 *PLoS One* **6** e26752
[10] Bochkarev V, Solovyev V and Wichmann S 2014 *J. R. Soc. Interface* **11** 20140841
[11] Pechenick E A, Danforth C M and Dodds P S 2015 arXiv:1503.03512
[12] Altmann E G and Gerlach M 2016 *Statistical Laws in Linguistics* (Berlin: Springer) pp 7–26
[13] Zipf G K 1936 *The Psycho-Biology of Language* (London: Routledge)
[14] Gerlach M and Altmann E G 2013 *Phys. Rev.* X **3** 021006
[15] Piantadosi S T 2014 *Psychonomic Bull. Rev.* **21** 1112–30
[16] Moreno-Sánchez I, Font-Clos F and Corral Á 2016 *PLoS One* **11** e0147073
[17] Pagel M, Atkinson Q D and Meade A 2007 *Nature* **449** 717–20
[18] Lieberman E, Michel J B, Jackson J, Tang T and Nowak M A 2007 *Nature* **449** 713–6
[19] Ferrer i Cancho R, Solé R V and Sol R V 2001 *J. Quant. Linguist.* **8** 165–73
[20] Petersen A M, Tenenbaum J N, Havlin S, Stanley H E and Perc M 2012 *Sci. Rep.* **2** 943
[21] Williams J R, Bagrow J P, Danforth C M and Dodds P S 2015 *Phys. Rev.* E **91** 5
[22] Wolfram MathWorld: Harmonic Number http://mathworld.wolfram.com/HarmonicNumber.html (Accessed: October 2016)
[23] Havrda J and Charvát F 1967 *Kybernetika* **3** 30–5
[24] Burbea J and Rao C 1982 *IEEE Trans. Inf. Theory* **28** 489–95
[25] Endres D and Schindelin J 2003 *IEEE Trans. Inf. Theory* **49** 1858–60
[26] Briët J and Harremoës P 2009 *Phys. Rev.* A **79** 052311
[27] Cover T and Thomas J 2006 *Elements of Information Theory* (New York: Wiley)
[28] Gerlach M, Font-Clos F and Altmann E G 2016 *Phys. Rev.* X **6** 021009
[29] Herdan G 1960 *Type-Token Mathematics* (Den Haag: Mouton)
[30] Heaps H 1978 *Information Retrieval* (New York: Academic)
[31] Gerlach M and Altmann E G 2014 *New J. Phys.* **16** 113010

J. Stat. Mech. (2017) 014002