Eur. Phys. J. B (2016) 89: 241
DOI: 10.1140/epjb/e2016-70423-9

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Power-law regularities in human language

Ali Mehri[a] and Sahar Mohammadpour Lashkari

Department of Physics, Faculty of Science, Noshirvani University of Technology, Babol, Iran

Received 8 July 2016 / Received in final form 31 August 2016
Published online 9 November 2016 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2016

Abstract. Complex structure of human language enables us to exchange very complicated information. This communication system obeys some common nonlinear statistical regularities. We investigate four important long-range features of human language. We perform our calculations for adopted works of seven famous litterateurs. Zipf's law and Heaps' law, which imply well-known power-law behaviors, are established in human language, showing a qualitative inverse relation with each other. Furthermore, the informational content associated with the words ordering, is measured by using an entropic metric. We also calculate fractal dimension of words in the text by using box counting method. The fractal dimension of each word, that is a positive value less than or equal to one, exhibits its spatial distribution in the text. Generally, we can claim that the Human language follows the mentioned power-law regularities. Power-law relations imply the existence of long-range correlations between the word types, to convey an especial idea.

1 Introduction

Many natural symbolic sequences such as language, music, genetic codes and neural signals are commonly applied for information conveyance. Human language is one of the important manifestations of natural languages, and its emergence could be regarded as a significant transition in hominid evolution [1]. It has special importance in human communication, culture, and even intelligence. Human beings employ language as an adaptive equipment to communicate and express their opinions. Words are constituents which interact with each other to form particular patterns. Such patterns represent human thoughts, feelings, will, and knowledge which are called meanings. The brain has a restricted capacity to store lexicons. On the other hand, human needs unlimited concepts to achieve a successful communication. The increasing need for new concepts is resolved by establishing complex syntactic and semantic relations between limited set of stored words and symbols. As a carrier of highly complex information, human language must operate under the competing requirements of allowing high information rate and at the same time being robust under communication errors [2]. The grammatical and semantic complexity of human language discriminates between mankind and other species. Uncovering the statistics and dynamics of human language helps in characterizing the universality, specificity and evolution of cultures [3].

A great deal of human knowledge has been included in the written part of language. Several universal features establish complexity of human written texts. Some of them like Zipf's law and Heaps' law refer to power-law relations related to word frequencies in text [4,5]. Besides, in statistical physics, entropy as a key concept can be applied to extract macroscopic properties of natural and artificial systems from their microscopic details. It could also be used as a quantitative measure of the (dis)order degree of a system, in the realm of information theory [6]. In quantitative linguistics, entropy suggested to measure the amount of information conveyed by the message [7]. In addition, fractal analysis is commonly used to measure the complexity of a system. Fractal dimension of the system or its multi-fractal spectrum can reveal its self-similar behavior [8]. Extraordinary spatial distribution of words in the text, makes it possible to convey a certain concept. The fractal analysis appears to be a good candidate for exploring the spatial patterns in texts, and finding their self-similar structures. The fractal dimensions of words can be easily computed by box counting method. It is shown that, the positions of a word type within the text form a fractal pattern with a specified dimension [9]. It should be expressed that, Linguistic laws are only meaningful if accompanied by a model for which the fluctuations can be computed [10].

We will analyze some famous literary texts, and we will evaluate the above mentioned important statistical laws of them. The nonlinear long-tail behaviors of the mentioned linguistic laws confirm the existence of long-range correlations between the words, to express a certain idea.

The organization of the remainder of the article is as follows. In Section 2, we briefly review Zipf's law, Heaps' law, information content and fractlity in the human languge. Then, we report our findings, in Section 3. The related graphs are illustrated, and the power-law exponents are extracted. Finally, in Section 4, we present a brief discussion and summary of the work.

[a] e-mail: alimehri@nit.ac.ir

2 Language statistics

In this section, we will briefly discuss four important statistical disciplines governing the human language. All of them follow power-law relations, that imply nonlinear behaviors in the language.

2.1 Zipf's law

George Zipf noted the manifestation of several robust power-law distributions arising in different realms of human activity. Among them, the most striking was the one referring to the word frequencies in human language [11,12]. It states that, if the most frequent word in a text is assigned rank 1, the second most frequent word is assigned rank 2 and so on, the frequency of any word decreases with its rank (rareness) in the text with a power-law decay [13]:

$$f \propto r^{-\zeta}. \tag{1}$$

In this equation, r is called the frequency-rank of a word, f is its frequency in a natural corpus and ζ is referred to as Zipf's exponent. Various aspects of the complexity of a communication system may depend on the value of the Zipf's exponent [14,15]. Zipf's law is probably the most intriguing and at the same time well studied experimental law of quantitative linguistics, and it is extremely popular in its wider sense in the science of complex systems.

A great number of systems in social science, economy, cognitive science, biology, and technology have been proposed to follow Zipf's law. All of them are composed of some elementary units, which are called tokens, and that these tokens can be grouped into larger, concrete or abstract entities, called types. Zipf's law deals with how tokens are distributed into types [16,17]. In the language case, types are the vocabularies which are distributed via a particular manner to express a special idea. For instance, texts can be divided into letters, morphemes, etc., but most studies in quantitative linguistics have considered the basic unit to be the word [18]. Zipf's law has been observed in many human languages, with different exponents depending on languages [19–21]. It is shown that, Zipf's plot has two scaling regimes with different exponents reect a division of word types into two lexical subsets, the kernel and unlimited lexica [22,23]. Bernhardsson et al. show that, the word-frequency distribution of a text, with a certain length written by a single author, has the same Zipf's exponent as a text of the same length extracted from an imaginary complete infinite corpus written by the same author [24]. Empirical analysis on the frequency of most popular keywords, indicates that their Zipf's exponent has completely different value in top journals from various subjects and in the journals of low impact factors [25].

2.2 Heaps' law

Heaps' law is an important empirical law in linguistics, which states that the vocabulary will keep growing with corpus size [5,26]. In other words, this law determines how the size of the employed index will scale with the size of the corpus. In the general case, Heaps' law implies a power-law relation between the number of distinct types and the collection size. This law says that the size of lexicon N_v (the number of particular words) in a text or set of texts of size N_t is determined by:

$$N_v \propto N_t^{\beta}, \tag{2}$$

where N_v is the number of distinct words when the text length is N_t, and $\beta \leqslant 1$ is the so-called Heaps' exponent. The Heaps' law may originate from the memory and productive nature of human language [27]. Heaps law can be considered as a derivative phenomenon if the system obeys the Zipf's law [28].

2.3 Entropy

Authors arrange the words in the text and spread them in a particular manner to convey their message. It is known that, words are distributed within N_t partitions (text size) in a particular routine to arise a specific concept. The grammatical rules determine where words should be placed within a sentence and specify the position of verbs, nouns, adverbs, and other parts of speech. These rules make short range correlations between the words in a sentence. On the other hand, a text imply an especial meaning using a specific arrangement (semantic ordering) of word types throughout the text. Hence, the long-range correlation can be seen between the positions of word types [29,30].

Let us assume, all occurrence positions of a given word, w, in the text with frequency $M(w)$ are included in the set, $T(w) = \{t_1, t_2, \ldots, t_M\}$. The spatial distance between two successive occurrences of a word type can be represented by $D(w) = \{d_1, d_2, \ldots, d_M\}$, where $d_i = t_{i+1} - t_i$. Head and tail of the text are connected (Fig. 1), so we will have $d_M = t_M - t_1$. One can define a spatial probability distribution for each word type which contains relative distances between its consecutive occurrences: $P(w) = \{p_1, p_2, \ldots, p_M\}$. $p_i = d_i/N_t$ can be interpreted as probability of existence a cycle with length d_i around the specific word type in the co-occurrence words network. The Shannon entropy associated with spatial distribution of word types, w, in the given text is defined as:

$$H(w) = \sum_{i=1}^{M} p_i(w) \log\left[p_i(w)^{-1}\right]. \tag{3}$$

Entropy measures the amount of information or uncertainty in a random variable concerning to a natural process. Zero value for entropy represents certain outcome for a random variable. On the other hand, if all outcomes are equally likely, entropy will have its maximal value. In our case, entropy shows degree of (dis)order in word usage pattern, and $H_{\max}(w) = \log(M)$ when a lexicon distributes homogeneously in the text.

The significant words, which are relevant to text subject, appear in specific parts of text to imply the considered idea ($H(\text{relevant}) \ll \log(M)$). On the contrary,

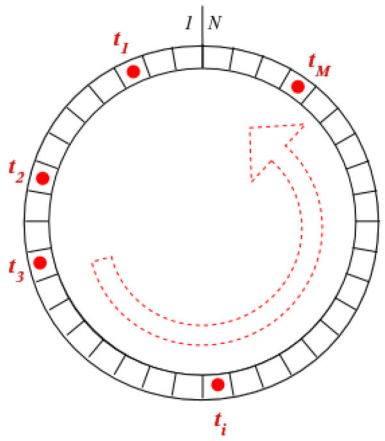

Fig. 1. Schematic diagram of cyclic model of text. Each cell represents position of a word in the text. Red circular bullets represent the occurrence positions of a particular word type. The arrow shows correct movement direction from head of the text to its tail [31].

the unimportant terms (i.e., articles, prepositions, conjunctions, etc.) are distributed almost homogeneously along the text by grammatical necessities. From physical point of view, it seems that in the meaningful natural and artificial texts, relevant words attract each other and tend to make clusters in ambient regions of the system. While the irrelevant words do not interact between themselves, and thus they appear randomly in the text since they do not feel each other. As a result, they set a nearly homogeneous spatial distribution: $H(\text{irrelevant}) \approx \log(M)$. Now, we can use an entropic measures, $\triangle H = |H(w) - \log(M)|$, to discriminate between content words and stop words in the given text. In this way, we sort all word types according to their $\triangle H$ value in descending order. Highly relevant words are positioned in the beginning of this sorted list [32].

2.4 Fractal dimension

A fractal is a rough or fragmented mathematical object that can be subdivided in parts, each of which is a reduced size copy of the whole. Fractal has the nature of self similarity in terms of statistics and fractal dimension is its important quantitative feature. The fractal dimension of a set is not equal to the topological dimension of space that the set is embedded in it [33]. It shows how detail of a fractal pattern changes with scale. A non-integral dimension can indicate the self similarity of the system.

Text is a certain arrangement of words in one dimensional array that carries a meaning. Any random shuffling of the words across the text significantly reduces its meaning; hence, the ordering of the words is important for representation of the meaning. In other words, the meaning shows a kind of regularity in pattern of occurrences of each word in textual array. If we consider the text array as a one dimensional space, the spatial pattern of occurrences of any vocabulary word will form a fractal pattern. Fractal dimension of a text can be defined by using words frequency and word length time series [34]. It can be obtained for sentence length variability [35].

We also can assign a fractal dimension to any word type in a given text, using the box counting method. In box counting, the text with length N_t is divided into N_l boxes with equal size l: $N_l = [N_t/l]$, where $[x]$ returns integer part of x. The number of boxes that contains a special word type (filled boxes) is denoted by $n_l(w)$ [9,32]. To calculate the fractal dimension of the selected word type, we should find the slope of the linear part of the box counting log-log plot,

$$\frac{n_l(w)}{N_l} \propto \left(\frac{l}{N_t}\right)^D. \tag{4}$$

Here D is referred to as fractal dimension of the selected word. Using this method, the fractal dimension of a word is generally $0 \leqslant D \leqslant 1$. When all occurrences of a word type are distributed uniformly across the text, all of the boxes have the same probability of containing a token of the word. Therefore, in this particular case, the number of filled boxes has the maximum possible value. In other cases, some of the boxes may contain more than one occurrence; this results in some of the other boxes remaining empty, and the number of filled boxes is less than this limiting value. The average fractal dimension of a given text is the average value of fractal dimension of its lexicons.

3 Results

We check the above mentioned laws for seven representative texts. The adopted texts are listed in Table 1. This is a linguistic notion that can be applied in many languages by delimiting sets of letters separated by spaces or punctuation marks. In this study, the texts were viewed as a sequence of 27 symbols, including the 26 letters of standard English and the space between words. Any other written symbols were disregarded. Zipf's law has been usually checked, by plotting the logarithm of the frequency versus the logarithm of the rank, and looking for some domain with a roughly linear behavior, with slope more or less close to –1. Figure 2 displays Zipf's plots for seven adopted books of well-known scholars. The vertical axis represents normalized frequency of word types, and the horizontal axis shows normalized rank. In order to calculate normalized frequency, we divide word frequency by text length. The normalized rank of each word type, can be obtained by dividing the rank by vocabulary size of the text. Zipf's exponents are obtained by using power-law fit to the empirical data extracted from above mentioned books. All fitting are performed with r-squares greater than 0.99. As seen in the figure, Zipf's exponent varies from $\zeta = 0.918$, in Weinberg's book, to $\zeta = 1.123$, in Cervantes' book.

The step-wise behavior of frequency for low ranks refers to hapax legomena. Traditionally, the words are divided into content and function words. The function words serve for establishing grammatical constructions, and their frequency directly depends on the sentence structure.

Table 1. Seven famous literary books, which are adopted to investigate power-law regularities. The text size, N_t, and the vocabulary size, N_v, of each book is also reported.

Book	Author	Date	N_t	N_v
Hamlet	William Shakespeare	1603	32 790	4576
Don Quixote	Miguel de Cervantes	1605	433 522	15 711
Moby Dick	Herman Melville	1851	221 564	17 148
The Origin of Species	Charles Darwin	1859	192 643	8180
The Evolution of Physics	Albert Einstein	1938	76 040	4561
The First Three Minutes	Steven Weinberg	1977	51 993	4033
The Theory of Everything	Stephen Hawking	2002	29 430	2795

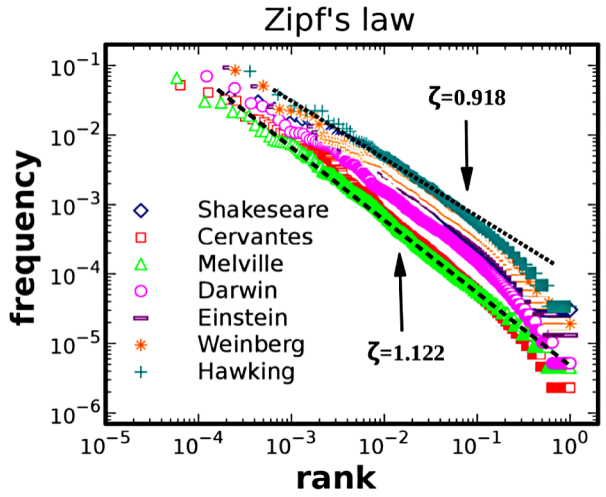

Fig. 2. Zipf's plot for adopted books of seven distinguished authors. The straight dotted and dashes lines correspond to power-law fits with Zipf's exponent $\zeta = 0.918$ and $\zeta = 1.123$, respectively.

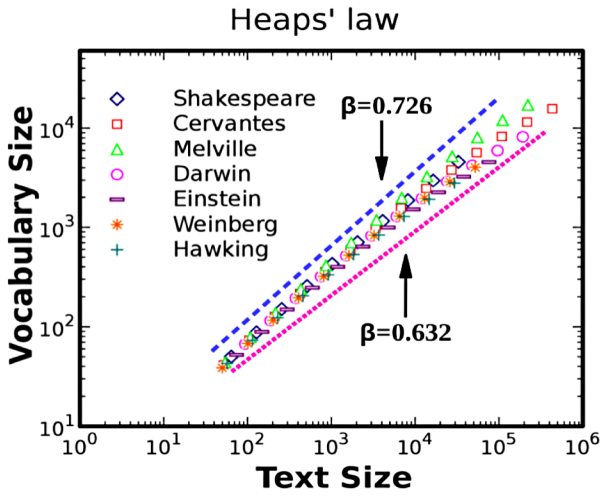

Fig. 3. Number of different words vs. length of given text for adopted books of seven famous authors. The straight dotted and dashed lines provide power-law fit to the extracted data. The non-linear growth of vocabularies with respect to text size, $0.632 \leqslant \beta \leqslant 0.726$, confirms Heaps' law in human language.

Since the percentage of function words is relatively high, the head of Zipf's plot contains mainly function words. But the majority of words in the central region of the plot have a high information content. A subset of those content words has a meaning that is specific for the text and can serve as its keywords [27,36]. It is worth mentioning that, maximum likelihood fitting can also be applied to extract Zipf's exponents [37].

We also check Heaps' law for above-mentioned texts (Tab. 1). In practice, the text is partitioned to P equal parts with size $n_t = N_t/P$. The mean vocabulary size of text partitions can be calculated as $n_v = \sum_{i=1}^{P} n_{v,i}/P$, where $n_{v,i}$ denotes vocabulary size of ith partition. We repeat this procedure for different partitioning sizes of the book to obtain the Heaps' plot. In Figure 3, we illustrate Heaps' plot for adopted works of seven famous writers. The boundary values of Heaps' exponent, in our adopted texts, are $\beta = 0.726$ for Melville's book and $\beta = 0.632$ for Darwin's book. Greater value of β exponent for Melville, indicates his far richer vocabulary treasury.

We illustrate the behavior of words' entropic measure, $\triangle H$, versus their normalized rank in Figure 4. It is clear that, all curves follow a Zipf-like regularity: $\triangle H \propto r^{-\alpha}$. The first part of entropy-rank graphs, with a small exponent, belongs to important words with significant information content. The tail of the graphs, with greater exponent, belongs to unimportant words with poor information content.

Mean entropy of a text, which is obtained by averaging over its all word types, yields the information content related to the spatial distribution of its words. The mean entropy associated with the mentioned representative texts, are reported in Table 2. Cervantes' book has the highest entropy, $H = 0.112$, among the adopted books. This high entropy value shows that Cervantes has distributed the applied word types in more homogeneous manner throughout his book, in comparison with the other books considered.

In Figure 5, we illustrate the relation between the normalized number of filled boxes, $n_l(w)/N_l$, and the normalized size of the boxes, l/N_t, for a typical word type (*sun*) in adopted works of seven well-known writers. As can be seen, all depicted graphs follow increasing power-law behavior for small box sizes. The exponent of this power-law regularity in known as the fractal dimension of the adopted word type. We report fractal dimension of above mentioned books in Table 2. The fractal dimension of text is defined as the average of the fractal dimension of its word types. The non-integral fractal dimension

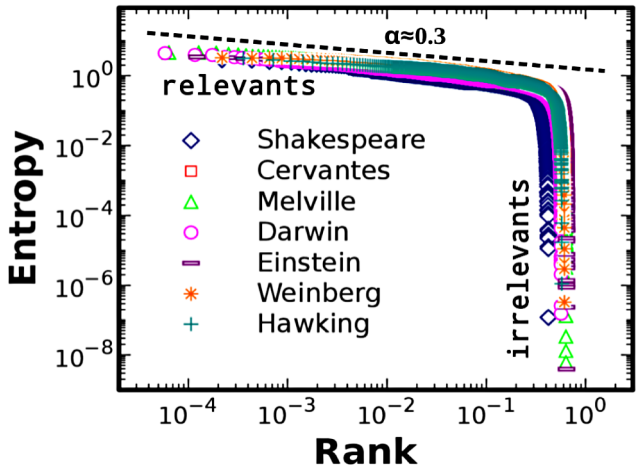

Fig. 4. $\triangle H$ vs. normalized rank (according to $\triangle H$) for adopted books of seven well-known writers as seven representative texts. All plots obey Zipf-like law with two different power-law exponents; a gentle slope region, with exponent $\alpha \approx 0.3$, for relevant words and a steep incline for irrelevant ones.

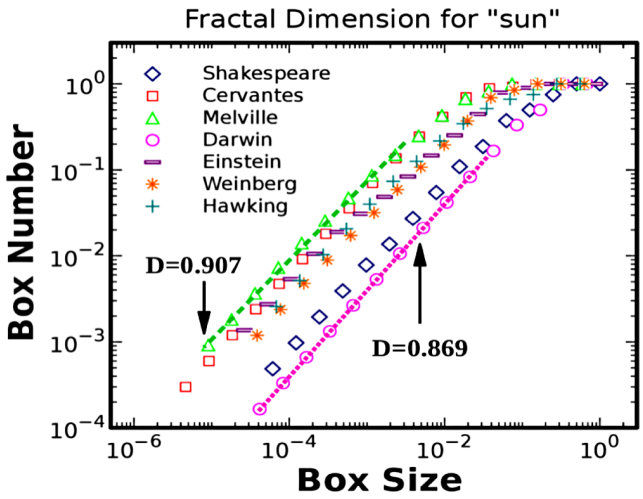

Fig. 5. Box counting results for a typical word type, in seven popular books considered in this study. Here, we depict normalized number of boxes that contain a special word type (vertical axis) vs. normalized box size (horizontal axis). The fractal dimension of the considered word (*sun*) changes from $D = 0.869$ in Darwin's and Einstein's works to $D = 0.919$ in Shakespeare's book.

confirms the existence of statistical self similarity in the human language.

4 Discussion and summary

In summary, in this article we addressed some important power-law aspects of human language. We focused on several statistical features, which confirm the presence of long-range correlations and complex relations in this language. Zipf's law, Heaps' law, information entropy and fractal structure of language is studied in this opportunity.

Table 2. Zipf's exponent (ζ), Heaps' exponent (β), mean entropy (H) and average fractal dimension (D) for representative works of seven outstanding writers. N_v and N_t denote vocabulary size and text size, respectively.

Author	ζ	β	H	D
Shakespeare	1.048	0.724	0.082	0.919
Cervantes	1.123	0.665	0.112	0.886
Melville	1.069	0.726	0.108	0.907
Darwin	0.989	0.632	0.096	0.869
Einstein	0.961	0.648	0.097	0.869
Weinberg	0.918	0.674	0.099	0.875
Hawking	0.933	0.671	0.091	0.875

We performed our calculations for representative works of seven outstanding authors. Our results are summarized in Table 2. All reported exponents are extracted by fitting the related data with power-law relations. High values of r-squared coefficient ($r^2 > 0.99$) indicate the goodness of the fits. In the case of Zipf's law we choose the first part of the plots, with smaller slope, for fitting process. We also perform normalization process to eliminate size effects in our results.

In the context of statistical linguistics, Zipf's law indicates a power-law relationship between the rank order of word types and the frequency of the appearance of them. The Zipf's law is the fundamental paradigm in statistical linguistics that serves as a prototype for rank-frequency relations and scaling laws in natural languages. Furthermore, the Heaps' exponent, β, quantify the vocabulary growth as a function of the text size.

We also assigned a quantitative measure to the information contained in the spatial distribution of words in the text, taking advantage of an entropic criterion. Mean entropy of a text is defined as the average of its word types' entropy. Moreover, we assumed that pattern of distribution of words in the text are fractal objects, and we introduced fractal dimension for them, using box counting method. One can obtain the average fractal dimension of a text, by taking average over fractal dimension of its all word types. The presented fractal dimension offers a significant clue into the spatial distribution of words, and the structure of language can be revealed by its mean value over all word types.

Power-law regularities confirm the existence of long-range correlations between words in texts under consideration, to express author's idea. Finally, this study shows that human language follows the fundamental statistical disciplines, as a manifestation of its linguistic complexity.

References

1. J.M. Smith, E. Száthmáry, *The Major Transitions in Evolution* (Oxford University Press, Oxford, 1997)
2. M.A. Montemurro, D.H. Zanette, arXiv:1503.01129v1 (2015)
3. L. Lü, Z.K. Zhang, T. Zhou, Sci. Rep. **3**, 1082 (2013)
4. G. Zipf, *Human Behavior and the Principle of Least Effort: An introduction to Human Ecology* (Addison-Wesley Press, Cambridge, 1949)

5. H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, New York, 1978)
6. T. Cover, J. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1991)
7. J. Sienkiewicz, M. Skowron, G. Paltoglou, J.A. Holyst, Advs. Complex Syst. **16**, 1350026 (2013)
8. M.F. Barnsley, *Fractals Everywhere*, 2nd edn. (Morgan Kaufmann, San Francisco, 1993)
9. E. Najafi, A.H. Darooneh, PLoS One **10**, e0130617 (2015)
10. E.G. Altmann, M. Gerlach, arXiv:1502.03296v1 (2015)
11. S.T. Piantadosi, Psychon. Bull. Rev. **21**, 1112 (2014)
12. D.H. Zanette, arXiv:1412.3336v1 (2014)
13. I. Moreno-Sánchez, F. Font-Clos, A. Corral, PLoS One **11**, e0147073 (2016)
14. J. Baixeries, B. Elvevåg, R. Ferrer-i-Cancho, PLoS One **8**(3), e53227 (2013)
15. X. Yan, P. Minnhagen, Physica A **444**, 828 (2016)
16. F. Font-Clos, A. Corral, Phys. Rev. Lett. **114**, 238701 (2015)
17. J.R. Williams, P.R. Lessard, S. Desu, E.M. Clark, J.P. Bagrow, C.M. Danforth, P.S. Dodds, Sci. Rep. **5**, 12209 (2015)
18. À. Corral, G. Boleda, R. Ferrer-i-Cancho, PLoS One **10**, e0129031 (2015)
19. A. Gelbukh, G. Sidorov, Lect. Notes Comput. Sci. **2004**, 332 (2001)
20. W. Deng, A.E. Allahverdyan, B. Li, Q.A. Wang, Eur. Phys. J. B **87**, 47 (2014)
21. A.M. Petersen, J.N. Tenenbaum, S. Havlin, H.E. Stanley, M. Perc, Sci. Rep. **2**, 943 (2012)
22. R. Ferrer i Cancho, R.V. Solé, J. Quant. Linguist. **8**, 165 (2001)
23. J.R. Williams, J.P. Bagrow, C.M. Danforth, P.S. Dodds, Phys. Rev. E **91**, 052811 (2015)
24. S. Bernhardsson, L.E. Correa da Rocha, P. Minnhagen, New J. Phys. **11**, 123015 (2009)
25. Z.K. Zhang, L. Lü, J.G. Liu, T. Zhou, Eur. Phys. J. B **66**, 557 (2008)
26. V.V. Bochkarev, E.Y. Lerner, A.V. Shevlyakova, J. Phys.: Conf. Ser. **490**, 012009 (2014)
27. M. Gerlach, E.G. Altmann, Phys. Rev. X **3**, 021006 (2013)
28. L. Lü, Z.K. Zhang, T. Zhou, PLoS One **5**, e14139 (2010)
29. M.A. Montemurro, D.H. Zanette, Adv. Complex Syst. **13**, 135 (2010)
30. A. Mehri, A.H. Darooneh, Physica A **390**, 3157 (2011)
31. A. Mehri, A.H. Darooneh, Phys. Rev. E **83**, 056106 (2011)
32. A. Mehri, M. Jamaati, H. Mehri, Phys. Lett. A **379**, 1627 (2015)
33. K. Falconer, *Fractal Geometry*, 2nd edn. (John Wiley & Sons, Chichester, 2003)
34. M. Ausloos, Phys. Rev. E **86**, 031108 (2012)
35. S. Drożdż et al., Information Sci. **331**, 32 (2016)
36. A.E. Allahverdyan, W. Deng, Q.A. Wang, Phys. Rev. E **88**, 062804 (2013)
37. A. Clauset, C.R. Shalizi, M.E.J. Newman, SIAM Rev. **51**, 661 (2009)