Word statistics in Blogs and RSS feeds

Lambiotte, R.; Ausloos, M.; Thelwall, M.

Published in:
Journal of Informetrics

DOI:
10.1016/j.joi.2007.07.001

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 1 (2007) 277–286
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Word statistics in Blogs and RSS feeds: Towards empirical universal evidence

R. Lambiotte [a,*], M. Ausloos [a], M. Thelwall [b]

[a] *GRAPES, Université de Liège, B5 Sart-Tilman, B-4000 Liège, Belgium*
[b] *SCIT, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK*

Received 22 February 2007; received in revised form 15 July 2007; accepted 16 July 2007

Abstract

We focus on the statistics of word occurrences and of the waiting times between such occurrences in Blogs. Due to the heterogeneity of words' frequencies, the empirical analysis is performed by studying classes of "frequently-equivalent" words, i.e. by grouping words depending on their frequencies. Two limiting cases are considered: the dilute limit, i.e. for those words that are used less than once a day, and the dense limit for frequent words. In both cases, extreme events occur more frequently than expected from the Poisson hypothesis. These deviations from Poisson statistics reveal non-trivial time correlations between events that are associated with bursts of activities. The distribution of waiting times is shown to behave like a stretched exponential and to have the same shape for different sets of words sharing a common frequency, thereby revealing universal features.

Keywords: Time statistics; Information networks; Zipf law; Activity pattern

1. Introduction

Web logs, also known as Blogs, have become an influential medium (Glance, Hurst, & Tomokiyo, 2004; Hammersley, 2005; Thelwall, Prabowo, & Fairclough, 2006), that encompasses a broad variety of subjects, e.g. politics and science, and are participative by nature. They involve a huge number of interacting users that belong to several layers of the population, from topic specialists to average people. This variety suggests that Blogs could be an efficient information source for identifying, tracking and modeling the spread of ideas and opinion formation, for example in public debates over political questions. Indeed, the democratic nature of Blogs allows us to examine how trends develop from the interactions of decentralized bloggers and to follow dynamic opinion changes over a wide and diverse sample of the population. This is in contrast with the main media where relatively few journalists are involved. Precise knowledge of word statistics in Blogs is consequently of interest in order to make coherent statistical tests for automatically detecting critical events, e.g. trends or media shocks (Kleinberg, 2002, 2008).

The most basic time statistics ignoring correlations between events can be modeled by Poisson distributions. This distribution concerns independent events: the number n of events arriving during some time interval Δ occurs with a

* Corresponding author.
E-mail address: Renaud.Lambiotte@ulg.ac.be (R. Lambiotte).

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2007.07.001

probability

$$P(n|a) = \frac{a^n}{n!}\mathrm{e}^{-a}, \tag{1}$$

where a is the arithmetic average number of events during this time interval. Moreover, the distribution of waiting times between two successive Poisson events is the negative exponential:

$$f(\tau) = \tau_c^{-1} \exp\left(-\frac{\tau}{\tau_c}\right), \tag{2}$$

where $\tau_c = \Delta/a$ is the average characteristic waiting time between events. This distribution is well-known to apply to nuclear disintegration but it has also been used for describing the time gaps between shoppers entering a store (Kan & Fu, 1997), the number of failure of products (Gregory, 2005), the number of terrorist acts (Telesca & Lovallo, 2006) as well as the number of airplane accidents as a function of time (Ausloos & Lambiotte, 2006a). An increasing amount of empirical evidence indicates, though, that human activity patterns do not fit this model. It has been shown by many other authors that human processes are rather heterogeneously distributed in time, with short periods of high activity (Kleinberg, 2002, 2008; Willinger & Paxson, 1998), or bursts, separated by long periods of inactivity (Barabási, 2005; Dewes, Wichmann, & Feldman, 2003; Dezsö et al., 2006; Ebeling & Neiman, 1995; Gopikrishnan, Plerou, Gabaix, Amaral, & Stanley, 2001; Paxson & Floyd, 1995; Sabatelli, Keating, Dudley, & Richmond, 2002; Vázquez, 2005; Vázquez et al., 2006). This heterogeneity is characterized by a distribution of waiting times which deviates from the exponential (2) and which, usually, presents a so-called heavy tail.

In this paper, we focus on the statistics of such waiting times between word occurrences in Blogs (and other similar periodically updated web sources) and also on the statistics of the number of word occurrences per day. To do so, we focus on texts published in 68022 RSS feeds during a period of 214 days and analyze two limiting cases. On one hand, we focus on very rare "events", namely words that occur on average less than once per day. It is shown that the frequency of words is very heterogeneous, so that the time statistics have to be measured in classes of "frequently-equivalent" words, i.e. words are discriminated through their total number of occurrences during the whole time period. This discrimination allows us to show that the distribution of waiting times deviates from the exponential (2), i.e. it is fitted by a stretched exponential and therefore presents an overpopulated tail. The deviation from the pure exponential is evaluated with the quantity ζ that measures the importance of the second moment of the time statistics. Interestingly, it is found that the shape of the distribution as well as the value of ζ do not depend on the class of words in which they are measured. On the other hand, we focus on events that occur many times per day on average. In that case, scaling laws are applied in order to smoothen the empirical results. Deviations from the Poisson statistics (1) are also found. Consequently, our results not only confirm that the dynamics of topics in Blogs present bursts of activity (Kleinberg, 2002, 2008; Willinger & Paxson, 1998) but they also provide tools in order to measure the importance of such bursts by comparing the empirical word statistics to a Poisson uncorrelated process.

2. Data description

2.1. RSS format

Really Simple Syndication (RSS) is an XML application designed to deliver brief summaries of the most recent updates of web sites (Hammersley, 2005), although it is flexible enough to incorporate other applications, such as reporting updates in digital libraries or search engine databases. Users with RSS reader software can subscribe to a range of RSS feeds based upon their interests, perhaps including favourite Blogs, some news sites or some special interest sites. The RSS reader will typically check each feed hourly and report to the user whenever new content is found. Each RSS feed contains a list of the most recent site updates, stored as separate XML items. When new content is added to the site, a new item will be added to the feed and the oldest one removed. Hence, when checking for updates, the RSS reader needs to parse feeds for items and report only items that are new, i.e. which were not in the feeds when they were previously checked.

RSS is also an attractive format for large scale data collection and analysis because it is typically concise and easy to parse text. In addition, its contents are easily time stamped so that time series can be generated. In contrast, web pages are typically much less concise and much harder to parse. Moreover, time series are difficult to generate from such

web pages because they typically reveal at best a last modified date (that is not automatically updated by the author). Like RSS, Blogs are more amenable to time series generation because each posting is dated and old postings are not normally modified.

2.2. Methodology

In the following, we focus on the data collected from 68022 RSS feeds from February 11th 2005 to October 2nd 2005. The list of feeds was obtained predominantly from Google, using its filetype:rss command, in conjunction with random mid-frequency words. The purpose of this method was to gain a wide range of types of sites supporting RSS feeds. A small proportion of the feeds, about 1%, were extracted from manual browsing of the web and the nowadays extinct completeRSS.comweb site. Altogether, the feeds are predominantly composed of Blogs, but also of other sources of online information, and they are mainly in English (estimated around 70–80%). At this point, it is also important to stress that the boundary between major news outlets and prominent Blogs has become blurred because the top bloggers have similar readerships as major newspapers. This difficulty justifies therefore our study of a heterogeneous collection, that encompasses several kinds of data sources, i.e. incorporating as well personal diary-like Blogs, professional specialist Blogs and newspaper RSS feeds. Let us also stress that one drastic event took place during the period under consideration, the London Attacks of July 7th 2005.

Recall that each text published by a blogger is called a post and is made of a sequence of words separated by punctuation, a blank space or markup (e.g. HTML or XML). The data collection was performed as follows. Every 24 h, all feeds were scanned and their content compared with the content observed in the last scan. All new posts are attributed to the new scanning time. Over the time period of 234 days, we observed 2,294,672 different words in the data set. In Fig. 1, we plot the number of posts containing a specific word as a function of its rank (the most frequently-occurring word has a rank 1, the second placed word has rank 2, . . .). Let us remark the deviations from the power-law, i.e. Zipf law $1/x^{\lambda}$ and from the Zipf–Mandelbrot law $1/(1+ax)^{\nu}$ (Montemurro, 2001), as those observed in (Ferrer-Cancho & Sole, 2001; Rousseau, 2002; van Raan, 2001). In contrast, the empirical curve of the presently examined data is very well fitted by a modified power-law of the form

$$\frac{1}{1+a_1x^{\gamma_1}+a_2x^{\gamma_2}}, \tag{3}$$

where $a_1 = 0.2, a_2 = 0.0004, \gamma_1 = 0.65$ and $\gamma_1 = 1.5$. Let us also stress that Eq. (3) includes two different characteristic exponents (Benguigui & Blumenfeld-Lieberthal, 2006; Martínez-Mekler, 2006) and that it is reminiscent of Tsallis-like distributions (Tsallis, 1988). The main point for the rest of this paper is that the rank function of Fig. 1 behaves *qualitatively* like a power-law, which implies that the distribution of the number of posts also behaves like a power-law (Adamic & Huberman, 2002). Consequently, this distribution is very wide and not peaked around its average value, i.e. the number of posts fluctuates enormously from one word to another word.

Before going further, one should also note that the above automatic scanning has been perturbed a few times due to technical problems, leading to gaps in time as seen in Fig. 2a. These missing scans have therefore led to the erroneous attribution of posts for the missing days and for the day that followed (see Fig. 2b). In order to perform a time analysis of word frequencies, we removed from the time series these anomalous data. After this data cleaning, there remained a 214 day time period. This cleaning does not change the shape of the curves of Fig. 1, but reduces the systematic errors bars for the following waiting times study.

3. Word statistics

3.1. Ensembles of equivalent words

Let us label each word by the index α. The number of posts in which this word occurs on day i is noted $W_{\alpha i}$. Moreover $W_\alpha = \sum_{i=1}^{214} W_{\alpha i}$ denotes the number of occurrences of α over the total time period. As discussed above, words may exhibit a large range of frequencies (1– 10^6). The spread of these frequencies may find its origin in many causes, e.g. the word “popularity” (two synonyms may be more or less popular) or “contextuality” (words associated to general and frequent contexts should be used more often). Such effects may be estimated by typing words in Google and counting the number of matches. For instance, synonyms like “clothes” and “garments” certainly have different

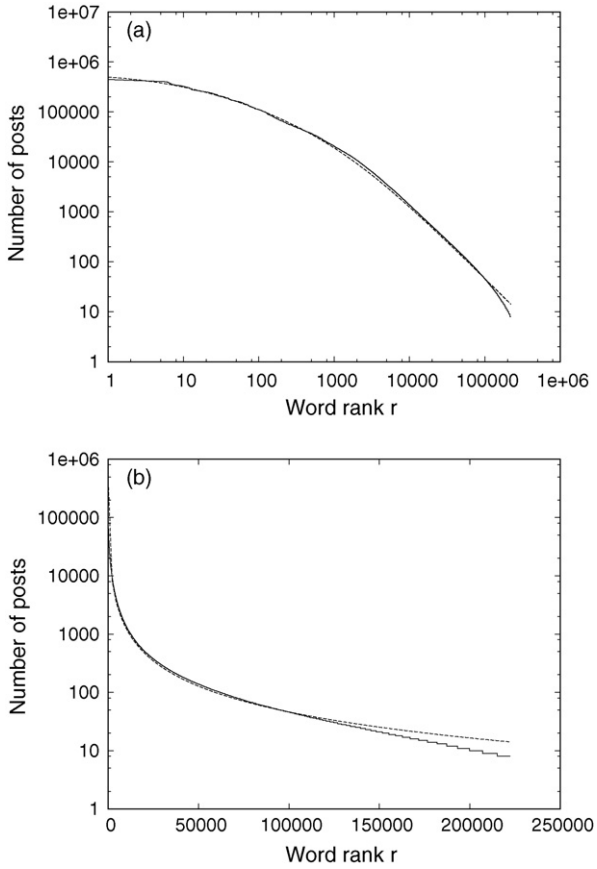

Fig. 1. Total number of posts containing a specific word as a function of its rank, in log-log scale (a) and log-normal scale (b). The first 10 most used words are, in decreasing order, (*the, a, to, of, and, in, for, is, on, it*). The curves show deviations from a power-law, but are quite well fitted (dashed line) by the modified power-law $\sim 1/(1 + 0.2x^{0.65} + 0.0004x^{1.5})$.

popularities, as their Google matches are 136×10^6 and 17×10^6, respectively. Similarly, a word associated with a popular topic/context, e.g. "music", which occurs 951×10^6 times, is used much more often than a word associated with a less popular topic, e.g. "tuberculosis" occurs 21×10^6 times.

It is well-known that heterogeneous events' frequencies may artificially overpopulate the tail of the distribution of waiting times. In (Ausloos & Lambiotte, 2006), for instance, it is shown that such an effect may even lead to a power-law distribution of waiting times while the system evolves in fact like a time-dependent Poisson process. This over-population originates from the fact that a distribution whose characteristic time fluctuates like

$$f(\tau) = \left\langle \tau_c^{-1} \exp\left(-\frac{\tau}{\tau_c}\right) \right\rangle_{\tau_c} \equiv \int d\tau_c \, \tau_c^{-1} \exp\left(-\frac{\tau}{\tau_c}\right) p(\tau_c), \tag{4}$$

where $p(\tau_c)$ is the probability that the characterstic time is τ_c, always exhibits larger fluctuations around the average waiting time than the Poisson distribution (2) does (Ausloos & Lambiotte, 2006; Beck, 2001). In order to overcome this difficulty, we separate out words depending on their frequencies. Define the ensemble E_k of words $\{\alpha_{i_1}, \ldots, \alpha_{i_{n_k}}\}$, that occur k times in the whole time interval. A word that is used only once is usually called a "hapax legomenon", while a word used twice is a "dis legomenon", thrice, a "tris legomenon", etc. Let us also denote the number of words belonging to the ensemble by n_k, i.e. it is the number of words α for which $W_\alpha = k$. In the following analysis, we consider that all words belonging to the same ensemble E_k are *a priori* equivalent. This assumption seems reasonable *a priori*, as words in the same ensemble have the same average waiting time and should be more homogeneous than

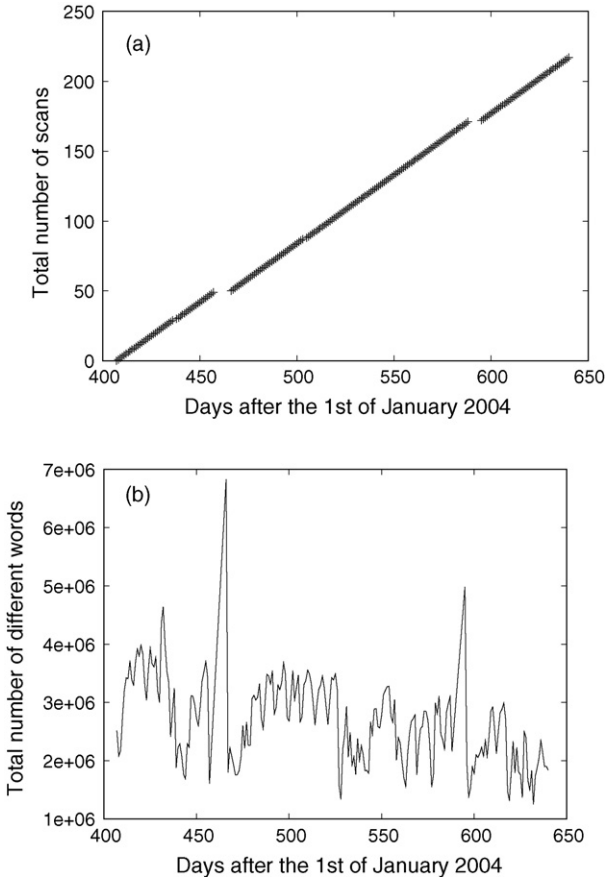

Fig. 2. In (a), time evolution of the number of performed scans. The discontinuities correspond to missed scans due to technical problems. In (b), we plot the measured number of different words as a function of time. Anomalous peaks are observed after each missed scan. These are removed from the analysis, as spurious data.

words randomly chosen in the whole set of used words. The validity of our assumption will be verified *a posteriori* by showing that waiting times are distributed in the same way in each ensemble E_k.

3.2. Dilute limit

Very rare words, i.e. words that occur much less than once a day on average, are ideal in order to test Poisson statistics, as the exponential distribution for waiting times (2) should fit them. Consequently, we focus in this section on the ensembles of words E_k, with $k < 214$, i.e. that occur on average less than once day. It is instructive to look at the distribution $f(\tau)$ (see Eq. (4)) obtained without splitting words into classes, i.e. by averaging the distribution over all words occurring $k < 214$ times. From the shape of that distribution (Fig. 3), one might conclude that time lags between word frequencies have a power-law distribution. We show in the following that this interpretation is erroneous and that the power-law shape is due to the averaging process described in the previous section. To do so, we measure the waiting time τ between two successive occurrences of one specific word in E_k, for each ensemble E_k separately. The distribution $f_k(\tau)$ is then obtained by performing the analysis for each word in E_k. It is shown (Fig. 4) that the width of f_k depends on the value of k (this is expected as each ensemble k is characterized by a different average frequency) and that f_k produces a *fat tail*, i.e. anomalously large probabilities for very large and very short time intervals. This fat tail suggests that word dynamics are dominated by bursts of activities (Vázquez et al., 2006) followed by long periods of rest in which the word does not appear. However, contrary to the distribution of Fig. 3, the distributions f_k are not well fitted by a power-law but resemble stretched exponentials

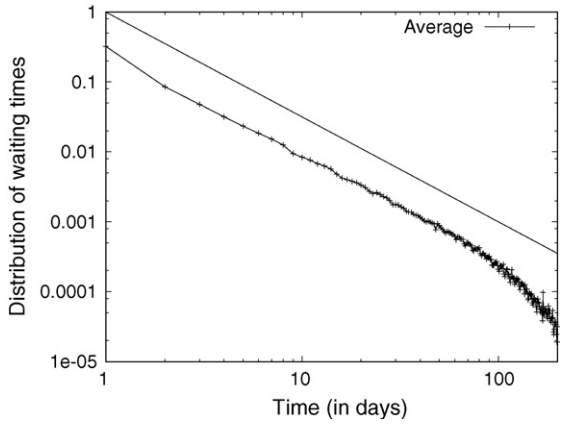

Fig. 3. Distribution of waiting times for all words in ensembles E_k, $k < 214$, as a function of time (in days), in a log-log scale. The solid line is the power-law $\tau^{-3/2}$.

$$f_k(\tau) = C\mathrm{e}^{-(a\tau)^{\nu}}, \tag{5}$$

where ν determines the shape of the distribution, C a constant of integration and a determines the time scale which all could be dependent on k. However, the exponent ν is always found to be very close to $\nu = 1/2$ for all the values of k. In that case, the constant of integration is $C = a/2$.

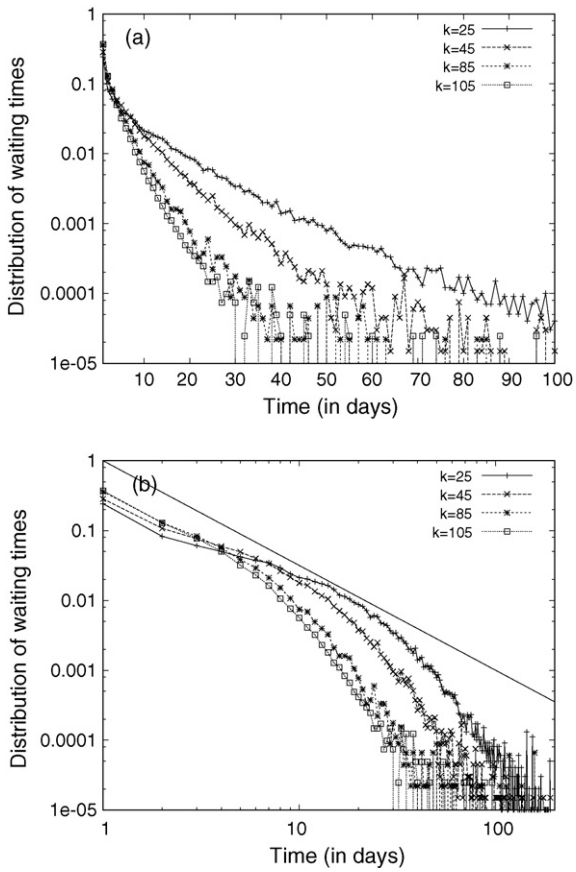

Fig. 4. Distribution of waiting times for four ensembles of words E_k, $k = [25, 45, 85, 105]$, i.e. belonging to the dilute limit case, as a function of time (in days): (a) in a log-normal scale and (b) in a log-log scale. The solid line is the power-law $\tau^{-3/2}$.

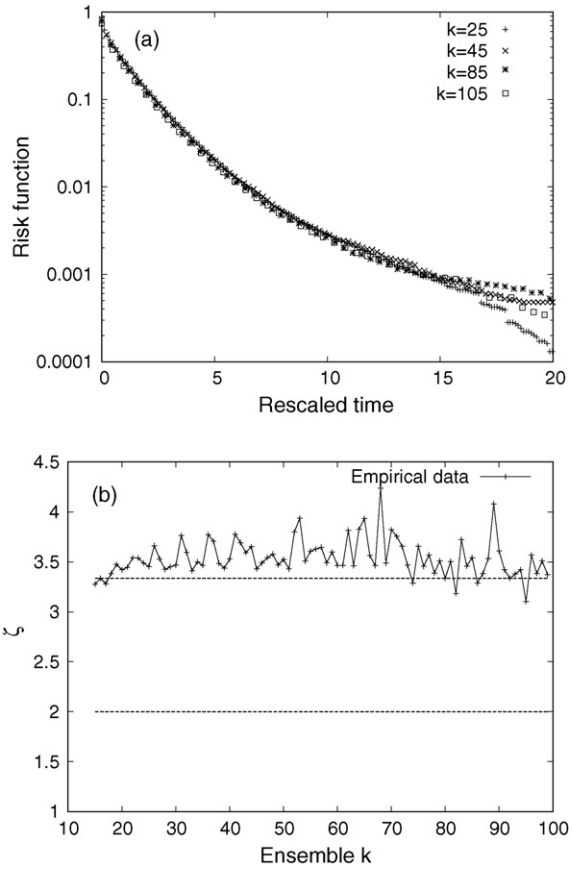

Fig. 5. In (a), empirical Risk function for four ensembles of words E_k, $k = [25, 45, 85, 105]$, as a function of the rescaled time t_R. See deviations from the exponential. In (b), we plot the quantity ζ as a function of the ensemble k in which it is measured (see text for the definitions of ζ and k). The dashed lines point to the value for a Poisson process, i.e. $\zeta = 2$, and for the stretched exponential (9), i.e. $\zeta = 10/3$.

Let us now show that the shape of the distributions $f_k(\tau)$ is universal. To do so, it is helpful to consider the Risk function $R_k(t)$

$$R_k(t) = \sum_{\tau=t}^{\infty} f_k(\tau) \tag{6}$$

in order to improve the statistics. The quantity $R_k(t)$ converges to zero for $t \to \infty$ in the same way the time distribution $f(\tau)$ does for the usual exponential and power-law time statistics (Ausloos & Lambiotte, 2006). By construction, words in the ensemble E_k are used k times in 214 days. Consequently, since the average waiting time $\langle\tau\rangle_k$ of such words is

$$\langle\tau\rangle_k = \sum \tau f_k(\tau) \sim \frac{214}{k}, \tag{7}$$

we change the time scale like $t \to t_R = t/(214/k)$. Empirical results for a large range of values of k as a function of t_R are shown in Fig. 5a and highlight deviations from the pure exponential, thereby confirming that correlations between word occurrences do not fit the Poisson hypothesis. Moreover, one observes that curves overlap for every k, thereby showing that the non-Poisson distributions are universal and that words belonging to different ensembles E_k share the same statistical properties. Note that this is observed over a large range of $k \in [25, 105]$.

In order to quantify the deviations from the exponential (2), it is useful to introduce the quantity

$$\zeta = \frac{\langle\tau^2\rangle}{\langle\tau\rangle^2}, \tag{8}$$

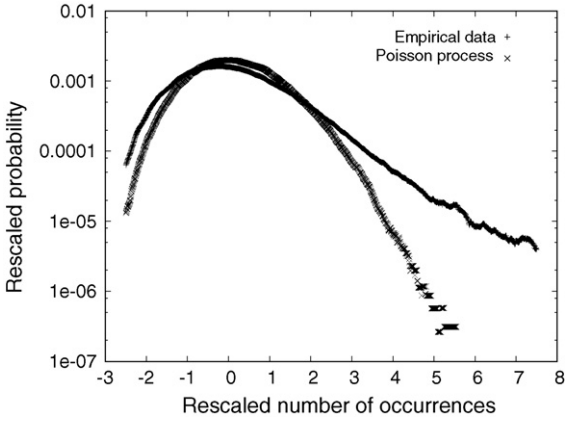

Fig. 6. Rescaled probability distribution of the rescaled number of occurrences Eq. (10) (+) that measures the deviations to the average $\langle x \rangle$. The data were obtained by averaging with the proper rescaling over all words occurring $k \in [1000, 2000]$ times, i.e. belonging to the dense limit case. This scheme has also been applied to Poisson random data numerically generated for the same values of k ($\times$).

where the average is performed over the distribution of waiting times. It is easily shown that $\zeta_{\text{poisson}} = 2$ when the process is Poisson, while it is larger than 2 if the fluctuations around the average waiting time are larger than those of a Poisson process. If the word occurrences were periodic, this quantity would go to 1. We have measured ζ for different ensembles E_k, $k \in [25, 105]$. It is shown in Fig. 5b that the empirical value is always larger than the Poisson value 2 and that it fluctuates around $\zeta_{\text{empirical}} = 3.5$. Interestingly, ζ_k does not depend on the ensemble E_k in which it is measured, which implies that the fluctuations around the average waiting time are the same in all ensembles and therefore confirms the universality observed in Fig. 5a. Let us stress that the empirical value $\zeta_{\text{empirical}} = 3.5$ is very close the value of ζ obtained from the observed distribution (5). Indeed, it is straightforward to show that $\langle \tau \rangle = 6/a$ and $\langle \tau^2 \rangle = 120/a^2$when

$$f_k(\tau) = \frac{a}{2} \mathrm{e}^{-(a\tau)^{1/2}}, \tag{9}$$

so that $\zeta = 10/3$ in that case.

3.3. Dense limit

For words occurring many times a day, it is rather meaningless to focus on the time lags between their occurrences, while a statistical analysis of the number of occurrences per day makes sense. In this limit ($k \gg 1$), however, the number n_k of words occurring k times is very low (see Fig. 1), so that a smoothing method is needed. Define $p(x, k)$ to be the probability that a word occurs x times one day, if it occurs k times over the 214 days. By definition, the average number of day occurrences is $\langle x \rangle = k/214$, but the width of the distribution is also expected to vary with k. From our data set, we can verify that the mean square displacement behaves like $\sigma \sim k^{-1/2}$, as expected. These two relations suggest to focus on the rescaled variable:

$$\tilde{x} = \frac{(x - (k/214))}{\sigma} \tag{10}$$

and on the corresponding rescaled probability distribution. By doing so, the data are smoothened and the characterization of the probability shape is possible.

In order to compare with Poisson events, we have generated numerically random ensembles E_k. This was done by randomly allocating k events into 214 boxes. The following step consists in measuring the distribution $p(x, k)$ and performing the above rescaling. As shown in Fig. 6, empirical data are less peaked around the average value, i.e. extreme events happen much more often than in the Poisson case. This over-representation leads to conclusions similar to those made in the previous section. In other words, even in the dense limit, bursts of activities occur.

4. Conclusion

In this article, we have performed an empirical analysis of the word frequencies arising in Blogs and RSS feeds. To do so, we have collected RSS data during a large time period (more than 200 days during spring 2005). These data encompass several kinds of information sources, such as newspaper RSS feeds and personal diary-like Blogs. Our analysis has been performed by discriminating words depending on their number of occurrences k. Namely, ensembles E_k of words occurring with the same frequency are defined and all words belonging to that ensemble are assumed to be "equivalent". This method is especially suitable when the frequency of word occurrences is very heterogeneous during the whole time window, as an heterogeneity of frequencies may radically alter the statistics of word occurrences.

Two limits have been considered: a dilute limit that consists of sparsely used words and a dense limit of words used many times a day. In the dilute limit, we have analyzed the statistics of waiting times between two successive occurrences of a word. It has been shown by using a proper rescaling that the distribution is the same for many ensembles E_k, thereby revealing a universal behaviour for word statistics. This universal distribution of waiting times has also been shown to deviate from the pure exponential, i.e. it behaves like a stretched exponential, and a statistical quantity ζ has been introduced in order to measure these deviations. Deviations from the Poisson distribution are also observed for the number of word occurrences per day in the dense limit. Altogether, these deviations are associated with *fat tails*, e.g. a high probability to observe extreme events, and suggest that word usage is dominated by bursts of activities followed by long periods of rest. Such bursts, which have also been observed in other social systems, e.g. Internet traffic (Willinger & Paxson, 1998), email or web browsing (Vázquez et al., 2006), may be caused by a response to an external triggering factor (e.g. US elections, publicity) (Lambiotte & Ausloos, 2006) or arise due to active endogenous discussions between bloggers.

Theoretical models reproducing the above empirical behaviour would be of great interest. Possible interesting ingredients include aging mechanisms (Cattuto, Loreto, & Servedio, 2006; Lambiotte, 2007; Wu & Huberman, 2007) that favour the realization of the most recent words as well as copying mechanisms in which people would have a tendency to use the words used by their acquaintances (Evans, 2007; Krapivsky & Redner, 2001; Lambiotte & Ausloos, 2005; Lambiotte & Ausloos, 2007).

Acknowledgement

This work has been supported by European Commission Project CREEN FP6-2003-NEST-Path-012864.

References

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the Internet. *Glottometrics*, *3*, 143.
Ausloos, M., & Lambiotte, R. (2006a). Time-evolving distribution of time lags between commercial airline disasters. *Physica A*, *362*, 513.
Ausloos, M., & Lambiotte, R. (2006b). A Brownian particle having a fluctuating mass. *Physical Review E*, *73*, 11105.
Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*, 207.
Beck, C. (2001). Dynamical foundations of non-extensive statistical mechanics. *Physical Review Letters*, *87*, 180601.
Benguigui, L., & Blumenfeld-Lieberthal, E. (2006). From lognormal distribution to power-law: A new classification of the size distribution. *International Journal Modern Physics C*, *17*, 1429.
Cattuto, C., Loreto, V., & Servedio, V. D. P. (2006). A Yule-Simon process with memory. *Europhysics Letters*, *76*, 208.
Dewes, C., Wichmann, A., & Feldman, A. (2003). An analysis of Internet chat systems. In *Proceedings of the 2003 ACM SIGCOMM conference on Internet measurement (IMC-03)* (pp. 51–64).
Dezsö, Z., Almaas, E., Lukács, A., Rácz, B., Szakadát, I., & Barabási, A.-L. (2006). Dynamics of information access on the web. *Physical Review E*, *73*, 066132.
Ebeling, W., & Neiman, A. (1995). Long-range correlations between letters and sentences in texts. *Physica A*, *215*, 233.
Evans, T. S. (2007). Exact solutions for network rewiring models. *The European Physical Journal B*, *56*, 65.
Ferrer-Cancho, R., & Sole, R. V. (2001). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, *8*, 165.
Glance, N. S., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. In *WWW 2004 workshop on the weblogging ecosystem: Aggregation, analysis and dynamics*
Gopikrishnan, P., Plerou, V., Gabaix, X., Amaral, L. A. N., & Stanley, H. E. (2001). Price fluctuations and market activity. *Physica A*, *299*, 137.
Gregory, P. (2005). *Bayesian logical data analysis for the physical sciences*. Cambridge U.P.
Hammersley, B. (2005). *Developing feeds with RSS and atom*. Sebastopol: O'Reilly.

Kan, K., & Fu, T.-T. (1997). Analysis of housewives' grocery shopping behavior in Taiwan: An application of the Poisson switching regression. *Journal of Agricultural & Applied Economics*, *29*, 397.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceeding of 8th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 91–101).

Kleinberg, J. (2008). Temporal dynamics of on-line information streams. In M. Garofalakis, J. Gehrke, & R. Rastogi (Eds.), *Data stream management: Processing high-speed data streams* (1st ed.). Springer, ISBN-10: 3540286071, ISBN-13: 978-3540286073.

Krapivsky, P. L., & Redner, S. (2001). Organization of growing random networks. *Physical Review E*, *63*, 066123.

Lambiotte, R. (2007). Activity Ageing in growing networks. *Journal of Statistical Mechanics: Theory and Experiment*, P02020.

Lambiotte, R., & Ausloos, M. (2005). Uncovering collective listening habits and music genres from collaborative networks. *Physical Review E*, *72*, 066107.

Lambiotte, R., & Ausloos, M. (2006). Endo- vs. exogenous shocks and relaxation rates in book and music 'sales'. *Physica A*, *362*, 485.

Lambiotte, R., & Ausloos, M. (2007). Growing network with *j*-redirection. *Europhysics Letters*, *77*, 58002.

Martínez-Mekler, G. (2006). Finite size universality in the Arts, Natural and Social Sciences. *Short communication at MEDYFINOL'06*.

Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A*, *300*, 567.

Paxson, V., & Floyd, S. (1995). Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions in Networking*, *3*, 226.

Rousseau, R. (2002). Lack of standardisation in informetric research. Comments on "Power laws of research output. Evidence for journals of economics" by Sutter M., & Kocher M. G. *Scientometrics*, *55*, 317.

Sabatelli, L., Keating, S., Dudley, J., & Richmond, P. (2002). Waiting time distribution in financial markets. *The European Physical Journal B*, *27*, 273.

Telesca, L., & Lovallo, M. (2006). Are global terrorist attacks time-correlated? *Physica A*, *362*, 480.

Thelwall, M., Prabowo, R., & Fairclough, R. (2006). Are raw RSS feeds suitable for broad issue scanning? A science concern case study. *Journal of the American Society for Information Science and Technology*, *57*, 1644–1654.

Tsallis, C. (1988). Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, *52*, 479.

van Raan, A. F. J. (2001). Two-step competition process leads to quasi power-law income distributions. Application to scientific and citation distributions. *Physica A*, *298*, 530.

Vázquez, A. (2005). Exact results for the Barabási model of human dynamics. *Physical Review Letters*, *95*, 248701.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K.-I., Kondor, I., & Barabási, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, *73*, 036127.

Willinger, W., & Paxson, V. (1998). Where mathematics meets the Internet. *Notices of the American Mathematical Society*, *45*, 961.

Wu, F., & Huberman, B. A. (2007). *Novelty and collective attention*. arXiv:0704.1158v1.