Proceedings of the 8th Polish Symposium of Physics in Economy and Social Sciences FENS, Rzeszów, November 4–6, 2015

Dynamics of Users Activity on Web-Blogs

Z. Kozioł*

National Center for Nuclear Research, Materials Research Laboratory, A. Sołtana 7, 05-400 Otwock–Świerk, Poland

Activity of users on Internet discussion forums is analyzed. The rank of users is shown to be approximated better by stretched-exponential function than by the Zipf law. Cumulative distribution function is found as an excellent tool in analysis of the dynamics of the collective social phenomena. We are able to approximate the number of blog comments with time by simple sigmoidal functions: probability of posting a comment is given by $P(t) = P_0(t)/[1 + P_0(t)]$, where $P_0(t) = (t/t_0)^\beta$ and β is close to 1.

DOI: 10.12693/APhysPolA.129.1060

PACS/topics: 89.65.–s, 89.20.Hh, 05.30.–d

1. Introduction

Collective social phenomena have been studied fruitfully in recent years with tools that have origin in statistical sciences [1, 2]. The subjects covered ranged from language, culture, opinion dynamics and spreading to crowd, behavior or hierarchy formation.

Here we address issue of users activity dynamics on discussion mailing lists, forums and blogs. Various aspects of the subject have been of broad interest recently [3–5]. Some of the regularities presented here are now known among researchers (in particular the relationship describing the number of entries to the mailing lists, or comments posted on blogs) on the rank of list member (its order in the list of the users that are most frequently writing there). We show that a stretched-exponential function is more suitable there than more commonly used the Zipf relation.

We show that activity in forums can be described with great accuracy by using a cumulative distribution function (CDF) of a sigmoidal form. The first observation of this, unpublished, was made in 1999 on mailing lists *IYP*†, *POLSKA*‡, *TLUG*§, *APAP*¶, and *POLAND-L*‖. Now new observations have been carried out on currently existing, active internet blogs, *Dziennik gajowego Maruchy* (*Marucha's blog*)**, *KiwiBlog*††, *My Nintendo News*‡‡, *Watts Up With That*§§, and *EU Times*¶¶.

The ubiquitous Zipf distribution, $P(x) \sim 1/x^{1+\mu}$, where μ is close to 1, describes well a broad range of phenomena. It arises naturally in structured, high-dimensional data [6]. It is characterized by scale invariance and by lack of scale. Not surprisingly, it has been considered even as a model useful in studies of some physical phenomena, in particular in statistical physics [7]. It is known also as a 80–20 rule (the Pareto law in economy): when income distribution is studied 80% of social wealth is found to be owned by 20% people [8, 9]. It is found in Internet traffic patterns; describes the number of pages on the web portals and the number of visits to web pages [10, 11], the terms searched most frequently by web users or results of some computer games [12], the intensity distribution of light or radio waves emitted by the galaxy [13], the distribution of citations of papers published [14], etc. It is argued that it arises in situations where we deal with random group division, where it predicts the existence of a unique group distribution with a power-law index determining the number of group elements and that index is in the range between 1 and 2, depending on the total size of the data set [15].

Alternatively, a stretched-exponential relation is used for studies of all these classes of phenomena, and in many cases it is found to describe them better than the Zipf law or its modified versions:

$$P(x) \sim exp(-x^\alpha). \tag{1}$$

Equation (1) origins from physics and it is most often

*e-mail: zbigniew.koziol@ncbj.gov.pl

†*IYP* (Internet Young Polonia Inc.) was a Polish-Canadian partisan organization led by this author, mainly of young Internet users from all over the world, especially students.

‡*POLSKA* discussion mailing list functioned actively for several years and was followed by the Marucha blog.

§*TLUG* (Toronto Linux Users Group) is one of the oldest, most influential and active community of users of Linux operating system. The list is not moderated.

¶*APAP* (Association of Polish American Professionals). A partisan organization / mailing list (with English as the language of discussion). Under moderate control.

‖*POLAND-L* was one of the most important Polonia mailing list, at the beginning of the wider use of the Internet.

**Marucha's blog exists since 2006. The blog is open for posting (comments) for all internet users. Spam and entries extremely controversial or vulgar tend to be rejected.

††KiwiBlog — general discussions, mostly on local issues, from New Zealand.

‡‡My Nintendo News — likely, a commercial PR site driven by Japanese company Nintendo Co., Ltd., with mostly young participants, game lovers.

§§Watts Up With That — an active community against restrictions on CO_2 emissions.

¶¶EU Times — an international newspaper based in Europe.

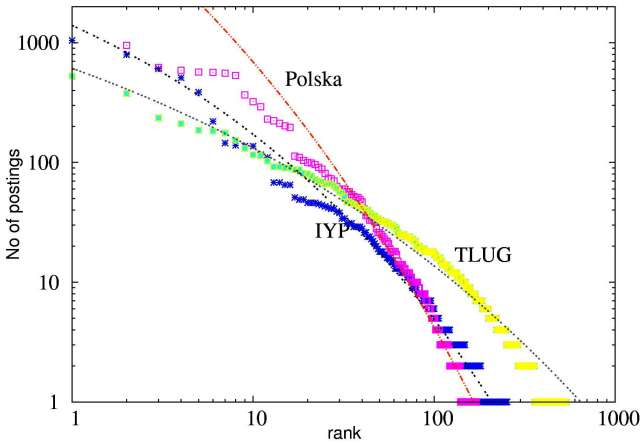

Fig. 1. Number of postings vs. users ranking on mailing lists *IYP*, *POLSKA* and *TLUG*. Fitting of power-law dependence on the tail side, with function $a/x^{1+\mu}$, would give exponent μ of about 0.9, 1.5, and 0.9, for *IYP*, *POLSKA* and *TLUG*, respectively. Fitting of function $aexp(-b \cdot x^{\alpha})$, is drawn by lines in the figure, with exponent α of 0.23, 0.19, and 0.17, for *IYP*, *POLSKA* and *TLUG*, respectively.

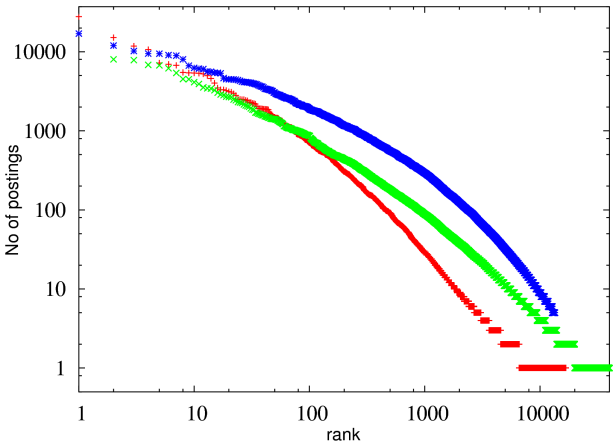

Fig. 2. Activity on blogs *EU Times*, *KiwiBlog*, *My Nintendo News*, *Marucha's blog*, and *Watts Up With That*, from left to right, respectively. The lines represent fitting of a stretched-exponential function, with parameters as given in Table I.

used to describe relaxation effects in disordered materials, as dielectric relaxation [16] or luminescence decays [17]. An overview of broad range of stretched-exponential distributions in nature and economy is given by Laherrere and Sornette [9].

2. The winner takes all

Results for mailing lists shown in Fig. 1 as well not shown data for lists *APAP* and *POLAND-L* are a good illustration of the 80–20 rule. During the studied period (between the beginning of 1997 and June 2000) there were 28510 messages sent to *POLAND-L*, and 25475 to the list *APAP*. It turns out that for both of these mailing lists just a few people dominated the discussions.

It is evident also that a stretched-exponential dependence, as shown by fitting lines in Fig. 1, is obeyed with a better accuracy than the Zipf law.

TABLE I

Fitting parameters of the data in Fig. 2. The exponent μ of equation $f(x) = x^{-(1+\mu)}$, obtained on tail side of the data, and exponent α of Eq. (1) are given.

Blog name	μ	α
Marucha's blog	0.67	0.155
KiwiBlog	1.5	0.22
My Nintendo News	0.45	0.13
Watts Up With That	0.82	0.145
EU Times	–0.18	0.13

Analysis of users activity on web blogs (Fig. 2) confirms that in the case of blog we deal with dependences like these for mailing lists. There is here the same pattern of activity, approximately described by power law relation (the Zipf distribution), and more exactly with the stretched-exponential function (1), as well dominance of a narrow group of users in frequency of posting comments.

The departure from the Zipf law is commonly observed in social networks [18] and it is treated as a signature of non-stationary growth of the social universe. There is an open question to what an extend we can compare exponents that describe size distribution of social groups and rank in discussion on mailing lists. It appears though that exponents found by us are close to these supposed to be "exact" in reference [18] ($\mu = 0.75 \pm 0.05$) while we have 0.9, 1.5 and 0.9, for *IYP*, *POLSKA* and *TLUG*, respectively (Fig. 1), and (this one must be more accurate) a value of 0.67 for the *Marucha's blog* (Fig. 2).

3. Stochastic dynamics of writing process

Dependences discussed so far do not say anything about the dynamics of the process of discussions on mailing lists or blogs. It seems that the approach we present here has not been used broadly so far, though some analysis in similar direction are known [19].

Graphs such as these in Fig. 3 were obtained by measuring the time interval between each successive entries on the blog. Then a function of the number of entries as a time interval between successive entries was created, and next the number of entries made has been normalized to unity at time tending to infinity, resulting in a CDF.

Intuitively, it is easy to interpret the meaning of CDF: the value of this function is the probability of the next entry being posted within time t. Hence, CDF describes the dynamics of blog posts/mailing list and as such is a characteristic function for a particular blog/mailing list.

Let us analyze "symmetry" of functions as these shown in Fig. 3. It turns out that nearly exactly the same curve

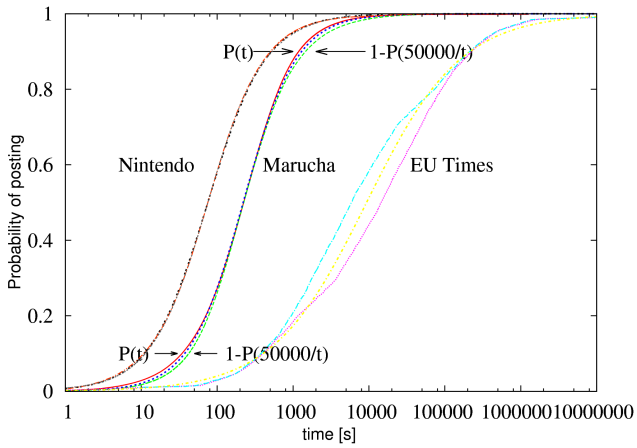

Fig. 3. Probability of posting $P(t)$ (cumulative distribution function, CDF) as a function of time on the *Marucha's blog*, *My Nintendo News* and *EU Times*. For each blog three datasets are shown, as explained in case of *Marucha's blog*: data for $P(t)$, transformed data $1-P(t_x/t)$, and between these two a line is shown computed with the function $P(t) = P_0(t)/(1.0+P_0(t))$, where $P_0(t) = (t/t_0)^\beta$. Parameters t_x, t_0 and β are listed in Table II.

as $P(t)$ is obtained when plot of the function $1-P(t_x/t)$ is made, where t_x is a certain parameter. A similar property has been also observed by us in the past for the data for mailing lists. That astonishing property leads us to assumption of the following approximation on $P(t)$:

$$P(t) = \frac{P_0(t)}{1.0+P_0(t)}, \tag{2}$$

where $P_0(t)$ is a monotonic function of t increasing from zero for small values of t to infinity at large values of t. Their simplest representation would be when $P_0(t)$ is assumed to be of power-law type. It turns out that such a relationship,

$$P_0(t) \sim (t/t_0)^\beta), \tag{3}$$

where t_0 is a fitting parameter, approximates well the data in Fig. 3. Equation (2) with $P_0(t)$ given by (3) and fulfilling the property $P(t) = 1-P(t_x/t)$ puts a restrain between allowed values of parameters t_0 and t_x: $t_x = t_0^2$. The data in Table II are in agreement with this requirement.

TABLE II

Parameters used for fitting CDFs in Fig. 3 to Eqs. (2), (3). The last column is the total number of postings.

Blog name	β	t_0	t_x	No.
My Nintendo News	1.14	75.3	5680	645000
Watts Up With That	1.34	88	7750	149000
Marucha's blog	1.24	220	50000	416000
KiwiBlog	1.05	74.6	5550	111000
EU Times	0.7	9300	85000000	2700

Initially, a blog popularity grows slowly, followed then by a qualitative change in activity (in case of *Marucha's blog* it is nearly parabolic growth). We are able to predict accurately the trends in the near future. The change of the users activity pattern at around $0.5-1.5\times10^8$ s in Fig. 4 (the year 2009/2010) is characteristic for all blogs studied.

It is interesting to answer the question whether the description of data in Fig. 3 is applicable in the case of discussions on narrow topics, under specific articles posted. To find the answer, we selected a few of the more active threads of high interest for a longer period of time, as described in Table III. It is found that activity in individual subjects is of the same nature as for the entire blog, except the parameters matching (β and t_0) this time are different. In particular, data in Table III suggest existence of a regularity: the smaller the exponent β, the larger the characteristic time t_0. To verify this, we analyzed the CDFs evolution as a function of time. The entire time-span studied (almost 8 years) has been divided to equal parts (of 5×10^6 s each) and t_0 and β were fitted to these data. Next, we were able to find that the following relation approximates well $t_0(\beta)$: $t_0 = 260000\exp(-5\beta)$. Hence, we are able to use a single equation that approximates the dynamics of activity through time-span of several years.

TABLE III

Parameters β and t_0 of Eq. (3) computed for a few discussion threads on *Marucha's blog*. The last column lists the total number of comments.

Start date	Subject (in Polish)	β	t_0	No.
2006/09/09	Neokatechumenat...	1.33	351	1896
2011/08/23	Pułapka na Rosję	1.1	899	722
2011/09/29	Wybory	0.95	1320	977
2010/04/25	Dariusz Kosiur...	0.88	5300	442

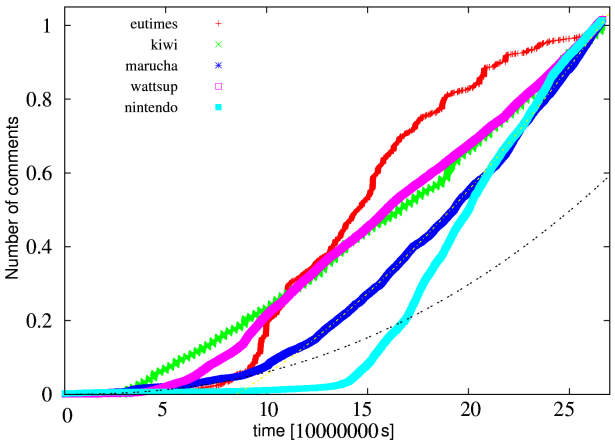

Fig. 4. Increase in number of postings as a function of time, for several blogs. The total number of postings has been normalized to 1 for every blog. In case of *Marucha's blog* the data are well described by the formula $f(x) = ax^b - c$, where $a = 10^{-14}$, $b = 2.3$ and $c = 0$ for short times and $a = 2.15\times10^{-9}$, $b = 1.704$ and $c = 70000$ for long times, as shown by solid lines.

4. Concluding remarks

The Zipf distribution describes well the number of entries from users of mailing lists and blogs as a function of their rank, on the side of large rank. An improved description is achieved when the stretched-exponential function is used instead.

The cumulative distribution function is found to be a good tool to study the dynamics of blog entries. Each mailing list has its own CDF. The results of the analysis suggest that the dynamics of entries in case of discussions on particular topic (thread) may be assigned by their own characteristic CDF.

For blogs or mailing list, the activity of all participants in the discussion, put together, can be accurately described using the function $P(t) = P_0(t)/[1 + P_0(t)]$, where $P_0(t) = (t/t_0)^\beta$, with β close to 1.

We are able to derive a single equation approximating the dynamics of activity on a web blog and predict its future development.

References

[1] C. Castellano, S. Fortunato, V. Loreto, *Rev. Mod. Phys.* **81**, 591 (2009).

[2] A. Pentland, *Social Physics. How Good Ideas Spread — The Lessons from a New Science*, The Penguin Press, New York 2014.

[3] P. Sobkowicz, A. Sobkowicz, *Europ. Phys. J. B* **73**, 633 (2010).

[4] N.K. Vitanov, M.R. Ausloos, *Models of Science Dynamics*, part of the series *Understanding Complex Systems*, Springer-Verlag, Berlin 2012.

[5] M. Suvakov, M. Mitrovic, V. Gligorijevic, B. Tadic, *J. R. Soc. Interface* **10**, 20120819 (2012).

[6] L. Aitchison, N. Corradi, P.E. Latham, arXiv:1407.7135 [q-bio.NC], 2014.

[7] K. Lukierska-Walasek, K. Topolski, *Mod. Phys. Lett. B* **28**, 1450088 (2014).

[8] M. Jagielski, R. Kutner, *Physica A* **392**, 2130 (2013).

[9] J. Laherrere, D. Sornette, *Eur. Phys. J. B* **2**, 525 (1998).

[10] S. Redner, *Eur. Phys. J. B* **4**, 131 (1998).

[11] L.A. Adamic, B.A. Huberman, *Glottometrics* **3**, 143 (2002).

[12] Z. Kozioł, http://www.nanophysics.pl/internet/glines.php.

[13] T. Mäkelä, A. Annila, *Phys. Life Rev.* **7**, 477 (2010).

[14] M.L. Wallace, V. Larivière, Y. Gingras, *J. Informetrics* **3**, 296 (2009).

[15] S.K. Baek, S. Bernhardsson, P. Minnhagen, *New J. Phys.* **13**, 043004 (2011).

[16] A. Milovanov, J.J. Rasmussen, K.R. Rypdal, *Phys. Lett. A* **372**, 2148 (2008).

[17] M.N. Berberan-Santos, E.N. Bodunov, B. Valeur, *Chem. Phys.* **315**, 171 (2005).

[18] Q. Zhang, D. Sornette, *Physica A* **390**, 4124 (2011).

[19] D. Rybski, S.V. Buldyrev, S. Havlin, F. Liljeros, H.A. Makse, *Proc. Natl. Acad. Sci.* **106**, 12620 (2009).