Dynamics of Users Activity on Web-Blogs

Zbigniew Kozioł

Research and Education Center of Microelectronics and Nanotechnology,
Rzeszów University, Rzeszów, Poland.

softquake@gmail.com, http://nanophysics.pl

- Activity of users on Internet discussion forums is analyzed.
- The rank of users can be approximated by stretched-exponential function.
- Cumulative distribution function is found as an excellent tool in analysis of the dynamics of this collective social activity.
- The number of blog comments with time can be approximated by functions that resemble Fermi-Dirac distribution function: $P(t) = P_0(t)/(1 + P_0(t))$, where $P_0(t) = exp(a \cdot log(t/t_0))$.

Marucha's blog (http://marucha.wordpress.com) exists since 2006. The blog is open for posting (comments) for all internet users. Daily a few new articles are added, and then commented by anonymous Internet users. Spam and entries extremely controversial or vulgar tend to be rejected.
IYP (Internet Young Polonia Inc.) was a Polish-Canadian partisan organization, mainly of young Internet users from all over the world, especially students. The mailing list had on average about 150 participants, with, through years, several thousands participants of discussions.
POLSKA discussion mailing list functioned actively for several years and was followed by the **Marucha's blog**.
TLUG (Toronto Linux Users Group) is one of the oldest, most influential and active community of users of Linux operating system. Talks are concentrated on technical aspects of using Linux, but are not limited to. 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 of administrator.
POLAND-L was one of the most important Polonia mailing list, at the beginning of the wide use of the Internet.
KiwiBlog (http://www.kiwiblog.co.nz) - general discussions, mostly on local issues, also politics, from New Zealand.
My Nintendo News (http://mynintendonews.com/) - most likely, a commercial PR site driven by Japanese company Nintendo Co., Ltd., with mostly young participants, game lovers.
Watts Up With That (http://wattsupwiththat.com/) - An active community from around the world against restrictions on CO_2 emissions.
EU Times (http://www.eutimes.net/) - a not very impressive discussion place by: *European Union Times is an international newspaper based in Europe with operational branches in America and Canada.*

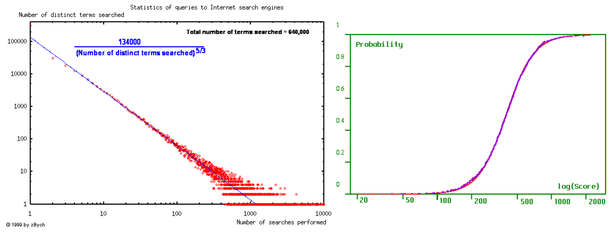
Introduction

Activity in forums can be described with great accuracy by using a function that resembles the Fermi-Dirac distribution, if cumulative distribution function (CDF) is used. Measurements are carried out on currently existing, active internet blog, *Dziennik Gajowego Maruchy* (*Marucha's blog*), *Kiwi Blog*, *EU Times*, *Watts Up With That* and *My Nintendo News*. Also results of analyses carried out for mailing lists *IYP*, *POLSKA*, *APAP* and *POLAND-L* are presented.

Zipf's versus stretched-exponential distributions

The ubiquitous Zipf's distribution, $P(x) \sim x^{-\mu}$, where μ is close to 1, describes well a very broad range of phenomena:

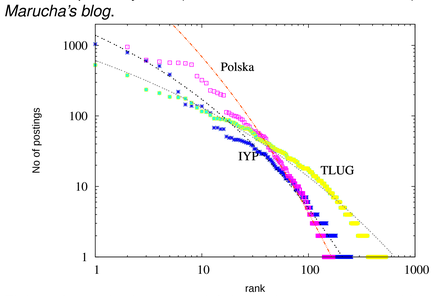
- Internet traffic patterns;
- The number of pages on web portals and the number of visits to web pages;
- The terms searched most frequently by web users; Results of computer games;
- The intensity distribution of light or radio waves emitted by the galaxy;
- The distribution of citations of papers published; Co-authorship popularity;
- The distribution of wealth in the population; Etc ...



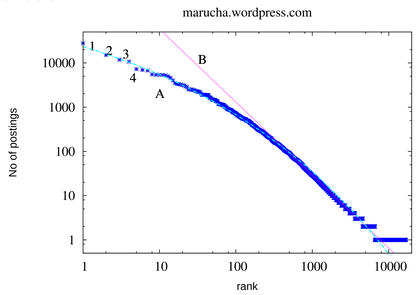
(Left Figure) Number of distinct terms searched as a function of the number of searches performed, for 640,000 distinct search terms. A simple power low dependence fits well to the data. (Right Figure) Probability distribution function of score (blue points) when playing glines. The red curve has been drawn according to $P(s) = s^a/(1 + s^a)$, where s=Score/k, with the following parameters: k=398, a=3.6.

The winner takes all.

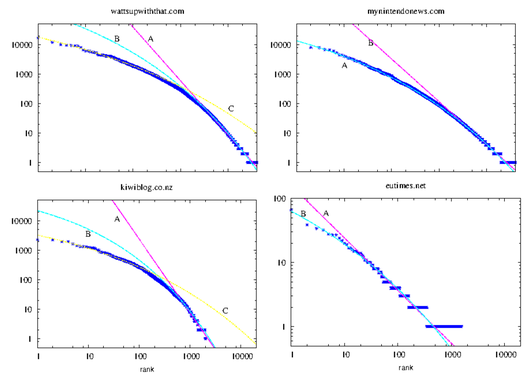
During the studied period (between the beginning of 1997 and June 2000) there were 28510 messages sent to *POLAND-L*, and 25475 to the list *APAP*. It turns out that for both of these mailing lists just a few people dominated the discussions. Here is a list of the most active users of the list Poland-L, with their number of postings (initials are given instead of real names): 1380 J.A., 1339 - W.G., 1225 - A.S., 924 - M.K., 784 - J.S.
The first two list users sent a total of about 10% of letters. And the first 5 sent about 20% of all letters. In contrast, 109 people sent a letter only once. During this time the number of participants exceeded slightly the number 300.
Analysis of users activity on *Marucha's blog* confirms that: in the case of blog we deal with dependencies like these for mailing lists.
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" (Zhang) ($\mu = 0.75 \pm 0.05$) while we have 0.9, 1.5 and 0.9, for *IYP*, *POLSKA* and *TLUG*, respectively, and (this one must be more accurate) a value of 0.67 for the *Marucha's blog*.



Fitting of stretched-exponential dependence of users ranking on mailing lists *IYP*, *POLSKA* and *TLUG*, with the function $a \cdot exp(-b \cdot x^{\beta})$, where exponent β is 0.23, 0.19, and 0.17, for *IYP*, *POLSKA* and *TLUG*, respectively, and a and b are certain numbers.



Activity on the *Marucha's blog* (points). Line B shows a simple power law relationship, $2300000 \cdot (x^{-1.67})$, and the line A a stretched-exponential dependence, $650000 \cdot exp(-3.4 \cdot x^{0.16})$. Numbers 1 to 4 refer to the firsts of the most active participants in the blog: 1 - Marucha, 2 - JO, 3 - Rysio, 4 - Bojkot166.



Activity on other blogs. Lines A show an attempt to fit a power law relationship, while lines B a stretched-exponential dependence.

Stochastic dynamics of writing process.

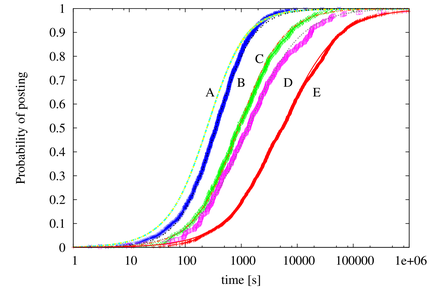
Dynamics of activity was analyzed by measuring the time interval between each successive entries on the blog. Then a function of the number of entries as a time interval between successive entries was created, and next the number of entries made has been normalized to unity at time tending to infinity, resulting in a cumulative distribution function (CDF).
Intuitively: the value of this function depending on the time t is the probability P of the next entry being posted within time t.
We start with the analysis of "symmetry" of the function P(t): Nearly exactly the same curve is obtained when plot of the function $1 - P(1/t)$ is made. That astonishing property leads us to assumption of the following approximation on $P(t)$:

$$P(t) = \frac{P_0(t)}{1.0 + P_0(t)}, \qquad (1)$$

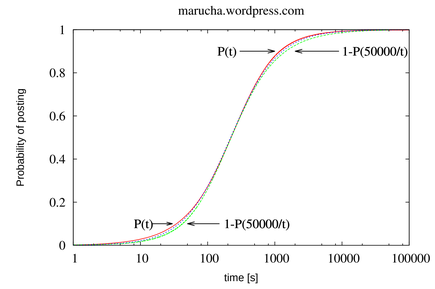
where $P_0(t)$ is a monotonic function of t increasing from zero for small values of t to infinity at large values of t. Additionally, we observe here that a function must be used of the kind where $P_0(t)$ has the property: $P_0(t) \sim 1/P_0(1/t)$ (that is easy to show by simple algebra). Their simplest representation would be that, when $P_0(t)$ is assumed to be of power-law type. Additionally, we should carry out appropriate normalization of t: it turns out that in fact, such a relationship,

$$P_0(t) \sim (t/t_0)^a), \qquad (2)$$

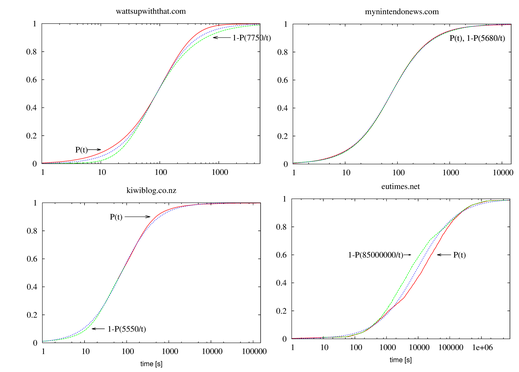
where a and t_0 are some fitting parameters, approximates well the data.
The function 2 is equivalent to the function in the form $exp(a \cdot log(t/t_0))$ - hence an analogy with the Fermi-Dirac distribution (FD), with the difference that in the case of FD exponent a is equal to 1. Here, the role of electrons (holes) energy in the solid state plays $log(t)$, and the Fermi potential role is played by the parameter $log(t_0)$.



Comparison of activity in a number of selected topics in the *Marucha's blog*. Line A represents the activity on the entire blog, and the remaining lines the activity in selected topics, as described in Table 1. For each data set a solid line is drawn described by the function $f(x) = f_0(x)/(1 + f_0(x)$, where $f_0(x) = exp(a \cdot log(x/t_0))$, and parameters a and t_0 are given in Table 1.



Probability of posting $P(t)$ (Cumulative Distribution Function) as a function of time on the *Marucha's blog*, marked with a red line. The green line represents a transform of $P(t)$ data in the form of function $1 - P(58000/t)$. The blue line (between the red and yellow one) shows the function $f(t) = P_0(t)/(1.0 + P_0(t)$, where $P_0(t) = exp(a \cdot log(t/t_0))$. The fitting parameters used were $t_0 = 244$ and $a = 1.22$.



Probability of posting $P(t)$ for other blogs. The fitting parameters used were $t_0 = 88$ and $a = 1.34$ for wattsupwiththat.com, $t_0 = 75.3$ and $a = 1.14$ for myntendonewes.com, $t_0 = 74.6$ and $a = 1.05$ for kiwiblog.co.nz and $t_0 = 9300$ and $a = 0.7$ for eutimes.net.

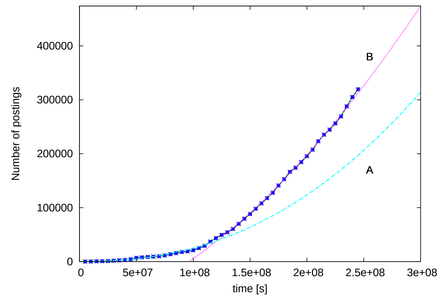
Long-time evolution.

Is sigmoidal description applicable in 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 1. Next figure shows the results observed from this kind of activity in individual subjects as well as for the entire blog, except that the parameters matching (a and t_0) this time are different.

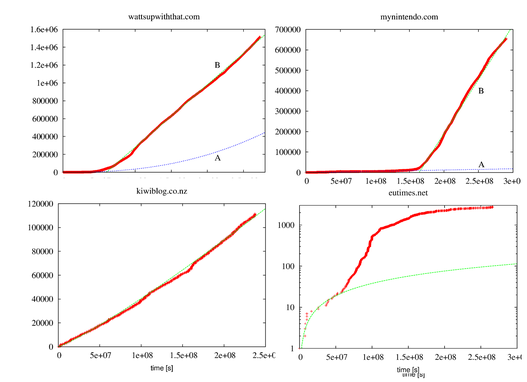
Table 1. Description of the data in the last Figure.

Line	Date	Subject discussed (in Polish)	a	t_0	Comments
A	all subjects		1.22	244	329929
B	2006/09/09	Neokatechumenat czyli kościół św. Kiko	1.33	351	1896
C	2011/08/23	Pułapka na Rosje	1.1	899	722
D	2011/09/29	Wybory	0.95	1320	977
E	2010/04/25	Dariusz Kosiur polski kandydat	0.88	5300	442

The data in Table 1 highlight the regularity: the smaller exponent a, the larger characteristic time t_0. In order to verify this and to describe more quantitatively the dynamics of users activity, we will analyze the CDFs for posting in narrower time range. The entire time-span studied (almost 8 years) has been divided to equal parts (of $5 \cdot 10^6$ s each, which is nearly 2 months). Next figure shows how the averaged number of postings changed with time: initially, when the blog was not widely known, its popularity grows slowly, and it is followed then by a steep nearly parabolic increase in number of postings. The change of the users activity pattern at around $1.2 \cdot 10^8 s$ (the year 2009/2010) is characteristic for many blogs, and reflects an intrinsic feature of users activity, after passing a certain critical value, and it requires verification by studies performed on other blogs.



Increase in number of postings as a function of time, measured in time intervals of $5 \cdot 10^6$ s. The solid lines fit well to activity on the blog, at the initial period of blog existence (line A) and line B at the later, current time. Both lines are drawn with the formula $f(x) = a \cdot x^b - c$, where $a = 10^{-14}$, $b = 2.3$ and $c = 0$ for line A and $a = 2.15 \cdot 10^{-9}$, $b = 1.7$ and $c = 80000$ for line B.



Increase in number of postings as a function of time, for several blogs.

Summary.

- Zipf's distribution describes well the number of entries from users of mailing lists and blogs as a function of their rank. An improved description is achieved when the stretched-exponential function is used instead.
- The cumulative distribution function as a function of time is found to be a good tool to study the dynamics of entries. Each mailing list has its own CDF function. The results of the analysis suggest that the dynamics of entries in case of discussions on particular topic (thread) may be assigned their own characteristic distribution function.
- 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 $P_0(t) = exp(a \cdot log(t/t_0))$. Similar relationship describes also the activity of the participants of discussions on specific topics.
- We are able to derive a single equation approximating the dynamics of activity on a web blog and predict its future development.

Selected References.

- Alex Pentland, *S*ocial Physics, The Penguin Press, New York, 2014.
- Claudio Castellano et al., *S*tatistical physics of social dynamics, http://arxiv.org/abs/0710.3256.
- Tad Hogg and Kristina Lerman, http://arxiv.org/pdf/0904.0016.pdf.
- Pawel Sobkowicz and Antoni Sobkowicz, http://arxiv.org/abs/1107.3275
- Zbigniew Kozioł, http://www.nanophysics.pl/internet/spy.php.
- Zbigniew Kozioł, http://www.nanophysics.pl/internet/glines.php.
- Maciej Jagielski and Ryszard Kutner, http://arxiv.org/abs/1301.2076.
- J. Laherrere and D. Sornette, Eur. Phys. J., 1998, B2, 525 – 539.
- Marcel Ausloos, http://arxiv.org/abs/1404.0269.