Eur. Phys. J. B (2012) 85: 28
DOI: 10.1140/epjb/e2011-20581-3

THE EUROPEAN PHYSICAL JOURNAL B

Regular Article

Neuropsychological constraints to human data production on a global scale

C. Gros[a], G. Kaczor, and D. Marković

Institute for Theoretical Physics, Goethe University Frankfurt, Germany

Received 18 July 2011 / Received in final form 17 October 2011
Published online 18 January 2012

Abstract. Which are the factors underlying human information production on a global level? In order to gain an insight into this question we study a corpus of 252–633 mil. publicly available data files on the Internet corresponding to an overall storage volume of 284–675 Terabytes. Analyzing the file size distribution for several distinct data types we find indications that the neuropsychological capacity of the human brain to process and record information may constitute the dominant limiting factor for the overall growth of globally stored information, with real-world economic constraints having only a negligible influence. This supposition draws support from the observation that the files size distributions follow a power law for data without a time component, like images, and a log-normal distribution for multimedia files, for which time is a defining qualia.

1 Author summary

The generation of new information is limited by two key factors, by the incurring economic costs and by the capacity of the human brain to process and store data and information; the controlling agent needs to retain an overall understanding even when data is generated by semi-automatic processes. These processes are reflected in the statistical properties of the data files publicly available on the Internet. Collecting a corpus of 252–633 mil. files we find that the statistics of the file size distribution are consistent with the supposition that data production on a global level is shaped and limited by the neuropsychological information processing capacity of the brain, with economic and hardware constraints having a negligible influence.

2 Introduction

Information production and storage becomes progressively easier. Moore's law [1] states that technological advancements lead to a doubling of computing power every 1.5 years and that data storage capacity increases by a factor of about 100 every 10 years [2]. Data production, which has the goal to increase knowledge and information, is constrained on one side by the economic costs involved and on the other side by the neuropsychological limitations and costs of the data generating agents. Maximizing the total amount of information generated for given amounts of economic and neuropsychological resources hence determines the shape of the file-size distribution [3].

[a] e-mail: gros07@itp.uni-frankfurt.de

The economic costs for data production involving hardware, software and management are proportional to the amount of data produced. The overall goal of data production is the generation of information, which can be measured by Shannon's information entropy [4]. Maximization of information entropy under the constraint of economic costs leads to file size distributions having exponential tails [3,5,6]. Exponential tails are however absent both in our data and in an earlier study of the file-size distribution on a large number of Windows file systems on desktop computers [7]. The absence of exponential tails for files hosted on Internet servers indicates that economic costs are not the limiting factors for data production.

The ability of the human brain to process and record information determines a subjective value which the producing individual attributes to an information source. E.g. the amount of information gained when increasing the resolution of a low quality image is substantially higher then when increasing the resolution of a high quality photo by the same degree. This relation is known as Weber-Fechner law and results from underlying neurophysiological processes [8–10]. We find that the observed file-size distributions on the Internet are consistent with the Weber-Fechner law and propose that neuropsychological constraints may be a dominant factor in shaping the statistics of global data production. This hypothesis is based on the finding that the distribution functions maximizing information entropy, given the neurophysiological constraints of the Weber-Fechner law, nicely reproduce the real world file-size distributions.

The neurophysiological constraints resulting from the Weber-Fechner law also imply that the different maximal

entropy distributions are qualitatively different for data formats involving time, like audio and video, compared to file types not involving time, as it is the case for images. We find that these distinct predictions are very well in agreement with the observed files-size distributions.

3 Maximal entropy distribution functions

Given a normalized distribution function $P(s)$ for a corpus of data, in our case the file-size distribution, its information content can be measured by Shannon's information entropy [4], $-\sum_s P(s)\log(P(s))$. The overall goal of data production, to attain an optimal information content, is achieved when the respective information entropy is maximal.

We denote with $c(s)$ the cost function associated with the economic and neurophysiological constrains. and with λ the respective Lagrange multiplier. The distribution function $P(s)$ maximizing information entropy given the constraint $c(s)$ is determined by [3,6]

$$\delta\Lambda[P] = 0, \quad \Lambda[P] = \sum_s P(s)\log(P(s)) - \lambda\sum_s P(s)c(s), \tag{1}$$

where $\delta\Lambda[P]$ denotes the variation of the functional $\Lambda[P]$ with respect to distribution functions $P(s)$. One obtains from (1) that $P(s) \sim \exp(-\lambda c(s))$. The maximal entropy distributions have then the form

$$P(s) \sim \exp\left[-\lambda_s s - \lambda_1 \log(s) - \lambda_2 \log^2(s)\right], \tag{2}$$

when considering cost functions containing terms proportional to the files size s, to $\log(s)$ and to $\log^2(s)$. The first term, linear in the size of the files s, corresponds to economic costs. The other two terms in the cost functions correspond to the scaling of neurophysiological resources.

The Weber-Fechner relations state that the neural representations of sensory stimuli [8], objects [9–11] and time perception [12] in the brain scale logarithmically with the intensity of the stimuli, the number of objects and the length of the time interval respectively. The perceived costs and benefits of information generation and processing hence scale logarithmically with physical data volume. Maximization of information entropy under the logarithmic cost function yields a power-law file size distribution, as described by equation (2).

The perceived cost function will scale furthermore as the square of the logarithm whenever the data is characterized by two neurophysiological distinct degrees of freedom, such as resolution and time. The distribution maximizing information entropy will then be a log-normal file size distribution, see equation (2). We find that this is indeed the case for multimedia files, such as audio and video files, for which the time is defining qualia. The file size distributions of non-temporal data types (e.g. texts and images) follow, on the other side, a power-law.

If the cost function scales as the square of the logarithm, the file-size distribution maximizing information entropy will then have a log-normal form, see equation (2).

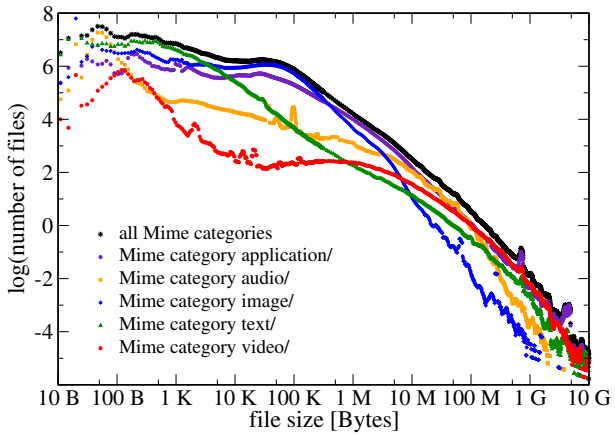

Fig. 1. (Color online) File-size distribution for 252 mil. files hosted in 7.7 domains. For all files types together and for the five Mime categories individually. Displayed is the $\log_{10}$ of the number of files in bins of 1 Kbyte.

We find that this is indeed the case for multimedia files, such as audio and video, for which the time is defining qualia. The file size distributions of non-temporal data types (e.g. texts and images) is closer, on the other side, to a power-law.

4 Results

We performed a large scale search of publicly available files on the Internet, utilizing the spider of file search engine FindFiles.net. For the corpus of hosts to be crawled we selected the collection of all outgoing links in Wikipedia.org and Dmoz.org, the open directory project, scanning in both cases all available editions. We crawled, in a first effort, a total of 7.7 mil. hosts, indexing 252 mil. data files. The resulting file size distribution is presented in Figure 1 in a log-log representation, spanning nine orders of magnitude. The crawling effort was then continued in a second stage until a corpus of 633 mil. files had been reached, which we used for a systematic study of the statistical properties of the resulting file-size distribution.

5 File taxonomy

Data files can be classified according to their Mime or Internet Media Types, e.g. a jpeg file has the Mime type *image/jpeg* within the Mime category *image/*. Five Mime categories make up about 99.9% of all data formats publicly accessible on the Internet, with *application/* contributing 33.2%, *audio/* 2.9%, *image/* 58.0%, *text/* 5.1% and *video/* 0.7% respectively to the total number of files in the Wikipedia/Dmoz corpus. The average number of files per host, the average file sizes (in Kbytes) and median file sizes (in Kbytes) are respectively (10.8|1312|136) for *application/*, (0.9|6733|1589) for *audio/*, (19.0|189|72) for *image/*, (1.7|3786|5) for *text/* and (0.2|28912|5548) for *video/*. The average file size of 189 Kbytes for images in

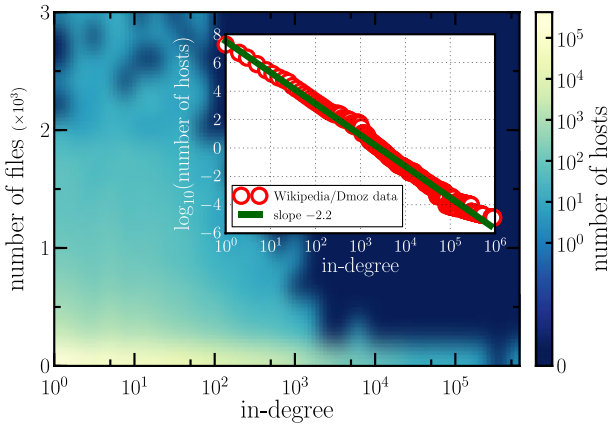

Fig. 2. (Color online) File hosting vs. in-degree. Main: the number of domains (dark blue: few hosts, white: many hosts) with a given in-degree (x-axis) and hosting a given number of files (y-axis); all Mime categories without text/ and image/. Inset: for the 32 mil. hosts receiving incoming links from the Wikipedia/Dmoz corpus the distribution of the in-degree.

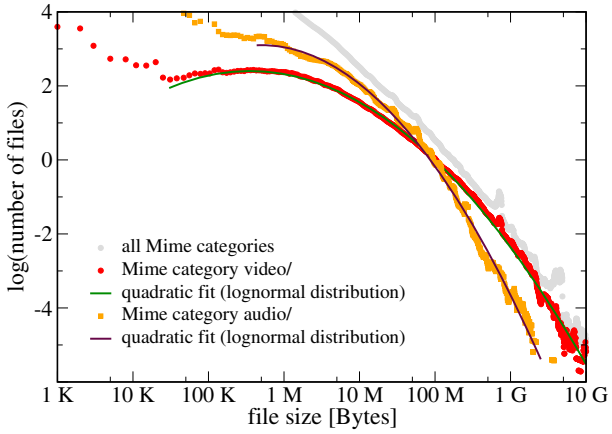

Fig. 3. (Color online) File-size distribution for videos and audio files (252 mil. files). Solid lines represent an eye guide of a quadratic form, $a\log(\text{size}) - b\log^2(\text{size})$, where $a, b > 0$.

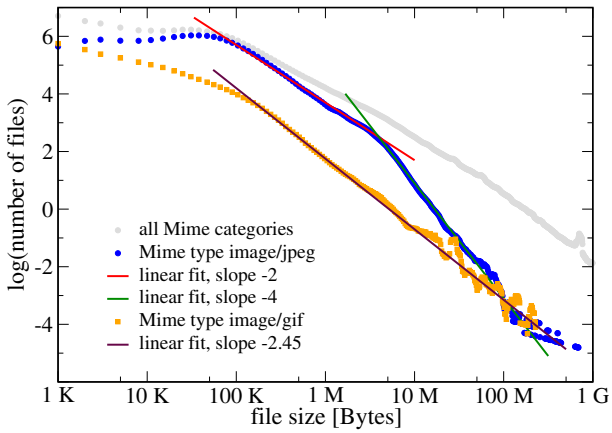

Fig. 4. (Color online) File-size distribution for jpeg and gif images (252 mil. files). The transition from a -2 to a -4 slope for the jpeg-file distribution occurs at about 4 Mbyte. This kink can be attributed to the transition from amateur to professional image production.

our data has seen an increase relative to the 15 Kbytes found in an earlier study [13]. The substantial difference between the respective means and medians is a consequence of the fat tails in the corresponding distribution functions, compare Figure 1.

6 File size distribution

Figure 2 shows the correlation between the number of files hosted and the in-degree (the number of inbound links) of the hosting domain. Important domains tend to have a high in-degree [14], e.g. the in-degree of Twitter.com is 805 000 in the Wikipedia/Dmoz corpus. The number of publicly accessible data files hosted is however anti-correlated with the in-degree, most data being hosted on relatively unknown hosts. The power-law for the in-degree distribution presented in the inset of Figure 2 has remained remarkably constant for the World Wide Web over the last decades. Our slope of -2.2 for the 32 mil. hosts within an one-click distance of the Wikipedia/Dmoz corpus is very close to the slopes between -1.94 and -2.1 found in previous studies [15–17].

A manifest property of the file size distribution presented in Figure 1 is the absence of exponential tails, which one would have expected [3,5] for an information entropy production constraint by economic limitations, like costs and available storage space. The lack of exponential tails has been observed in an earlier study of the file size distribution on a large number of Windows file systems on desktop computers [7]. They have also found that the utilization ratio of desktop hard disks is, on the average, below the capacity. Thus, the full storage volume is rarely utilized by the average PC user.

A differentiated perspective can be obtained when examining the functional form of the file size distributions for distinct Mime categories and types. The tails for the video and audio file distributions, shown in Figure 3, and the tails for the file size distributions of jpeg and gif images presented in Figure 4 differ manifestly.

The linear dependence observed in Figure 4 corresponds to a scale-free power-law $P(s) \propto s^{-\gamma}$ of the file size distribution $P(s)$ with distinct slopes for lossless and lossy image compression formats, gif and jpeg, respectively. For video and audio files the file size distribution follows a log-normal dependence, with $\log(P(s)) \propto \alpha\log(s) - \beta\log^2(s)$ fitting the data excellently over more than 5 orders of magnitude. These two distributions differ qualitatively in two aspects, namely in the occurrence of the quadratic term $\log^2(s)$ for the log-normal distribution and in the sign of the linear term. The leading term $-\gamma\log(s)$ has a negative slope for image data formats and a positive slope $\alpha\log(s)$ for the *audio/* and *video/* Mime categories (with $\alpha, \gamma > 0$). The log-normal dependence observed for video and audio files is hence qualitatively distinct with respect to a power-law scaling and cannot be interpreted

as a quadratic correction to a linear fit within a log-log data analysis.

The fact that the file size distributions and the distribution tails are qualitatively different for multimedia and image file formats, strongly indicates that they are determined by the underlying neurophysiological limitations of the producing agents. The cost functions are therefore, see equation (2), proportional to $\log(s)$ and $\log^2(s)$ for data characterized by one and two degrees of freedom, respectively.

7 Fitting methods

The guide-to-the-eye fits shown in Figures 3 and 4 indicate that the statistical properties of the file-size distributions depend on the presence/absence of a time-component.

For a systematic analysis we used the corpus of 633 mil. files, containing four Mime categories, *image/* (64.8%), *application/* (31%), *audio/* (3.5%) and *video/* (0.7%). The evaluation of file size distribution was performed in two steps. In a first step the files were binned into 1 Kbyte bins. In a second step we evaluated maximum likelihood estimates for two model distributions [18].

We analyzed the tails of the respective file size distributions with two types of discrete probability distributions, a power law,

$$p(k) = \frac{1}{Z_{k_{min},\alpha}} k^{-\alpha}, \qquad Z_{k_{min},\alpha} = \sum_{k_{min}}^{\infty} k^{-\alpha},$$

and one having a log-normal form

$$p(k) = \frac{1}{Z_{k_{min},\mu,\sigma}} \frac{1}{k} e^{-\frac{(\log k-\mu)^2}{2\sigma^2}},$$

$$Z_{k_{min},\mu,\sigma} = \sum_{k_{min}}^{\infty} \frac{1}{k} e^{-\frac{(\log k-\mu)^2}{2\sigma^2}}.$$

In both cases a lower bound k_{min} is introduced as a free parameter, as we don't expect describing the whole range of data but only the tails of available data by either a power-law or log-normal distribution.

The actual fitting procedure consist of following steps:

- We've first performed a maximum likelihood estimate for the lower bound k_{min} in the range from 1 KB up to 100 MB.
- Then, we have selected a k_{min} which minimizes residual sum of squares (*rss*) of the differences between the empirical and the fitted tails of the complementary cumulative distribution functions, that is

$$rss = \sum_{k'=k_{min}}^{k_{max}} \left(\Pr\left(k \geq k'\right) - F(k')\right)^2, \qquad (3)$$

with $F(k) = \sum_{k'=k}^{\infty} p(k')$ being the complementary cumulative distribution of the model and and $\Pr(k \geq k')$ respectively the empirical complementary cumulative file distribution.

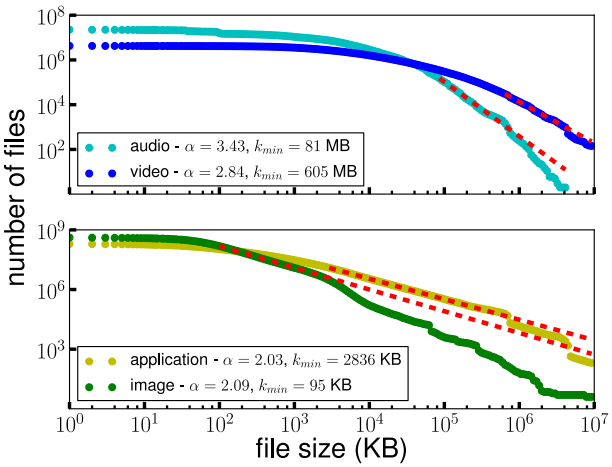

Fig. 5. (Color online) Power-law fit to the complementary cumulative file-size distribution. The red dashed lines are the fits, the respective parameters are given in the insets. For Mime categories *audio/*, *video/* (top) having a time component and *application/*, *image/* (bottom) having no time component.

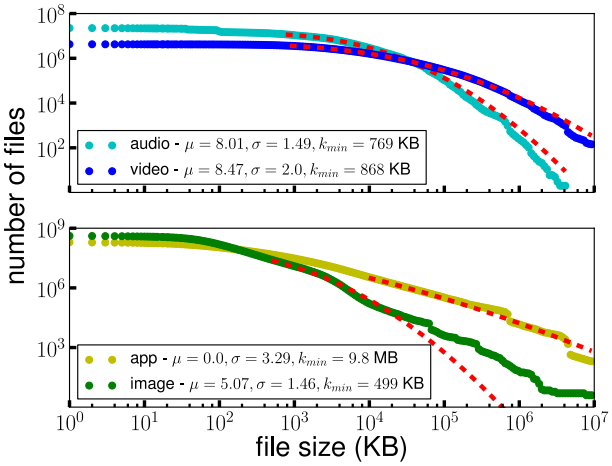

Fig. 6. (Color online) Log-normal fit to the complementary cumulative file-size distribution. The red dashed lines are the fits, the respective parameters are given in the insets. For Mime categories *audio/*, *video/* (top) having a time component and *application/*, *image/* (bottom) having no time component.

We present the best fits for the four Mime categories (*image/*, *application/*, *audio/* and *video/*) for a power-law distribution in Figure 5, and a log-normal distribution in Figure 6, respectively. In obtaining the maximum likelihood estimate for model parameters we have neglected files larger then 10 Gbytes, as there are only very few of these extremely big files, they are hence statistically not representative.

Comparing the two fits for *audio/* and *video/* data we find that the log-normal distribution describes the empirical data substantially better. The *rss* values are order of magnitude lower in the case of log-normal fit (see Tab. 1). The log-normal fit is also able to describe a broader range of the data then the power-law fit (compare Figs. 5 and 6).

Table 1. Residual sum of squares, *rss*, estimated as sum of square differences between empirical and fitted complementary cumulative distributions (see Eq. (3)).

	power-law fit	log-normal fit
image/	0.7	1.0
application/	3.3	47.9
audio/	26.4	2.1
video/	451.4	2.7

Similarly, a power-law fit for *application/* file-size distribution describes a broader range of the empirical data, and has an order of magnitude smaller value, then the one obtained for a log-normal fit (Tab. 1). In the case of the *image* file types the evidence in favor of a power-law distribution is not particularly strong, a consequence of the kink at around 4 Mbytes, compare Figure 4. Both fits, log-normal and power-law, describe a similar data range and the corresponding *rss* values are of similar magnitude.

8 Discussion

For images the production costs are functionally dependent on one variable, the resolution, which defines, modulo compression algorithms, the file size. The cost function for the production of videos depends however on two distinct quantities, the resolution per frame and the total number of frames, viz the time needed to shoot the sequence. Analogously for audio files, with frequency resolution and length being the two cost defining quantities. The cost functions associated with information production are hence one- and two-dimensional for images and audio/video formats respectively. We generically observe in our data that one-dimensional cost functions result in power-law file size distributions, two-dimensional cost functions in log-normal distributions.

For compound Mime categories or types, like *text/*, superpositions of these two basic distributions are observed. This correlation between dimensionality of data type and resulting file size distribution, which can be seen in Figures 5 and 6, finds a straightforward rationale when accounting for the neuropsychological constraints for data processing.

Our analysis is based on the assumption that an ensemble average over many information producing agents reveals the underlying information theoretical principles driving data production on a global level. Other studies have investigated alternative approaches, like the study of microscopic models capable of generating distributions with large tails, such as scale-free [19,20] and log-normal [21,22] and the double Pareto-lognormal distribution [23]. In a related context a log-normal distribution has been found for the distribution of city sized and be related to proporionate growth mechanisms [24–26].

Both approaches, the modelling of generative processes and the information theoretical perspective, are complementary and do not exclude each other. Ultimately it may be possible to derive classes of microscopic generative models from comprehensive information theoretical principles, as it has been proposed, e.g., for intrinsic neural adaption rules generating information entropy maximizing firing rate distributions [5,27].

We thank the file search engine http://www.findfiles.net FindFiles.net for support and data collection. The complete raw data of the Wikipedia/Dmoz corpus is available for download at http://www.findfiles.net/public

References

1. G.E. Moore, Proc. IEEE **86**, 82 (1998)
2. J. Gray, P. Shenoy, Proc. International Conference on Data Engineering **16**, 3 (2000)
3. C. Gros, *Complex and adaptive dynamical systems: A primer*, 2nd edn. (Springer Verlag, 2010), pp. 92–93
4. C.E. Shannon, W. Weaver, *The Mathematical Theory of Information* (University of Illinois Press, Urbana, 1949)
5. D. Marković, C. Gros, Phys. Rev. Lett. **105**, 068702 (2010)
6. B. Mandelbrot, An information theory of the statistical structure of languages, in *Communication Theory*, edited by W. Jakcson (Betterworth, 1953), pp. 486–502
7. J.R. Douceur, W.J. Bolosky, Perform. Eval. Rev. **27**, 59 (1999)
8. S. Hecht, J. General Physiology **7**, 235 (1924)
9. A. Nieder, E.K. Miller, Neuron **37**, 149 (2003)
10. S. Dehaene, Trends Cogn. Sci. **7**, 145 (2003)
11. A. Nieder, D.J. Freedman, E.K. Miller, Science **297**, 1708 (2002)
12. T. Takahashi, Med. Hypoth. **65**, 691 (2005)
13. S. Lawrence, C.L. Giles, Nature **400**, 107 (1999)
14. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.U. Hwang, Phys. Rep. **424**, 175 (2006)
15. R. Albert, H. Jeong, A.-L. Barabási, Nature **401**, 130 (1999)
16. L.A. Adamic, B.A. Huberman, Science **287**, 2115a (2000)
17. J.S. Kong, N. Sarshar, V.P. Roychowdhury, Proc. Natl. Acad. Sci. **105**, 13724 (2008)
18. A. Clauset, C.S. Shalizi, M.E.J. Newman, SIAM Rev. **51**, 661 (2009)
19. A.L. Barabási, R. Albert, Science **286**, 509 (1999)
20. X. Gabaix, P. Gopikrishnan, V. Plerou, H.E. Stanley, Nature **423**, 267 (2003)
21. M. Mitzenmacher, Internet Mathematics **1**, 226 (2004)
22. M. Mitzenmacher, Internet Mathematics **1**, 305 (2004)
23. W.J. Reed, M. Jorgensen, Commun. Stat. Theory Methods **33**, 1733 (2004)
24. J. Eeckhout, Am. Econ. Rev. **94**, 1429 (2004)
25. M. Levy, Am. Econ. Rev. **99**, 1672 (2009)
26. J. Eeckhout, Am. Econ. Rev. **99**, 1676 (2009)
27. J. Triesch, Neural Comput. **19**, 885 (2007)