Frequency of Occurrence of Numbers in the World Wide Web

Sergey N. Dorogovtsev, José Fernando F. Mendes, João Gama Oliveira
Department of Physics, University of Aveiro, 3810-193 Aveiro, Portugal

The distribution of numbers in human documents is determined by a variety of diverse natural and human factors, whose relative significance can be evaluated by studying the numbers' frequency of occurrence. Although it has been studied since the 1880's [1, 2], this subject remains poorly understood. Here, we obtain the detailed statistics of numbers in the World Wide Web, finding that their distribution is a heavy-tailed dependence which splits in a set of power-law ones. In particular, we find that the frequency of numbers associated to western calendar years shows an uneven behavior: 2004 represents a 'singular critical' point, appearing with a strikingly high frequency; as we move away from it, the decreasing frequency allows us to compare the amounts of existing information on the past and on the future. Moreover, while powers of ten occur extremely often, allowing us to obtain statistics up to the huge 10^{127}, 'non-round' numbers occur in a much more limited range, the variations of their frequencies being dramatically different from standard statistical fluctuations [3, 4]. These findings provide a view of the array of numbers used by humans as a highly non-equilibrium and inhomogeneous system, and shed a new light on an issue that, once fully investigated, could lead to a better understanding of many sociological and psychological phenomena.

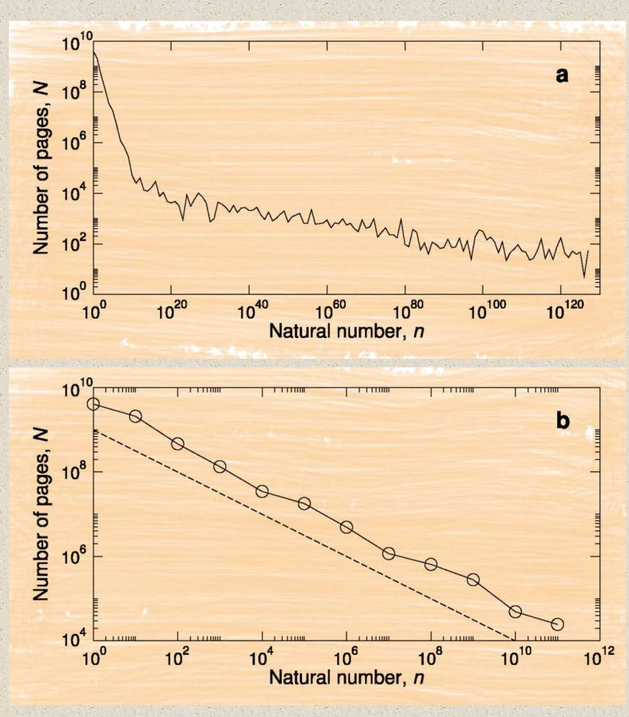

Figure 3: The frequencies of WWW pages containing powers of 10. **a,** The full log-log plot up to the maximal searchable 10^{127} (Google allows searching for strings with size up to 128 characters). **b,** The power-law-like part of the distribution, blown up from **a**. The dashed line has slope -0.5, therefore the decay is well approximated by

$$N(n) \sim 1/\sqrt{n}\,.$$

We emphasize that the power-law dependence is observed over 11 orders of magnitude, which is a uniquely wide range. Note that the crossover seen in **a** turns out to be close to the maximum 32 digit binary number.

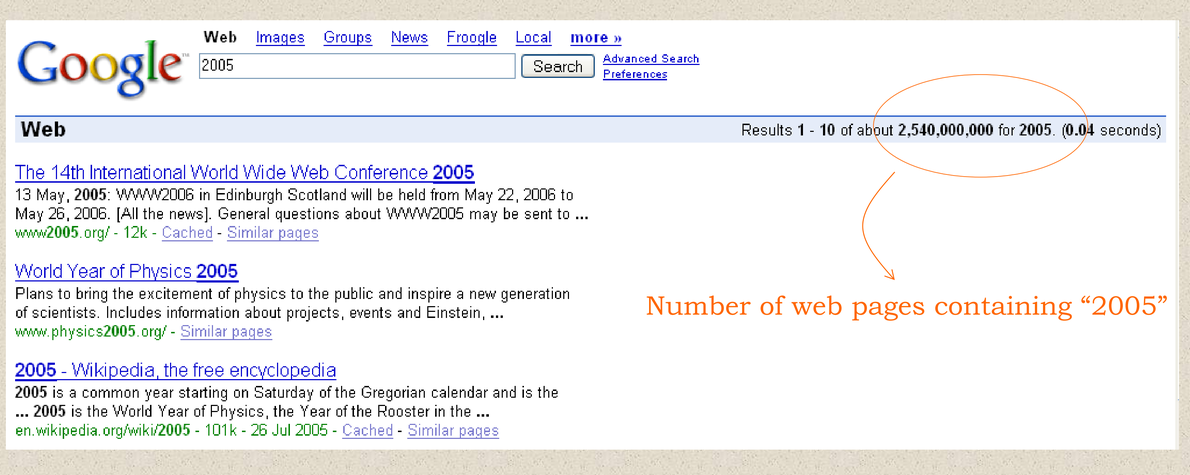

Figure 1: The occurrence frequency of "2005" as reported by Google in July 2005.

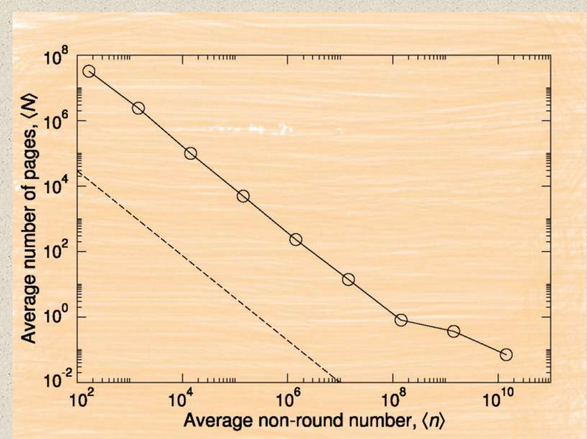

Figure 4: Log-log plot of the frequencies of web pages holding 'non-round' numbers. The circles show the average number of web pages with non-round numbers taken from relatively narrow intervals (50 numbers). Each interval is centered at the $\langle n \rangle$ coordinate of a circle. The dashed line has slope -1.3. Note that the power-law behavior is observed over 6 orders of magnitude. Non-round numbers occur much less frequently than powers of 10, which explains the essentially narrower range of numbers in this plot than in Fig. 3a.

TABLE I: **Typical numbers with high frequencies of occurrence**

Example	Description
1000	powers of 10
2460, 2465	'round' numbers: multiples of 10 and 5
666,[a] 131313	numbers easy to remember or symmetric
$512 = 2^9$	powers of 2
666,[a] 777	numbers with strong associations
78701	popular zip codes
866, 877	toll free telephone numbers
1812	important historical dates
747, 8086	serial numbers of popular products
314159	beginning parts of mathematical constants

[a]A number may occur simultaneously in several lines of the table.

Table I

With the help of Google search engine, we measure the frequency N of natural numbers (i.e. positive integers) n occurring in the World Wide Web (Fig. 1 and 2a). Our findings show that there are 'special' numbers which occur more often (Table I). In particular the frequency of numbers associated to years has a critical singularity at the current year (Fig. 2b). As we move to the past the frequency decays as a power law with exponent -1.2 (inset of Fig. 2b).

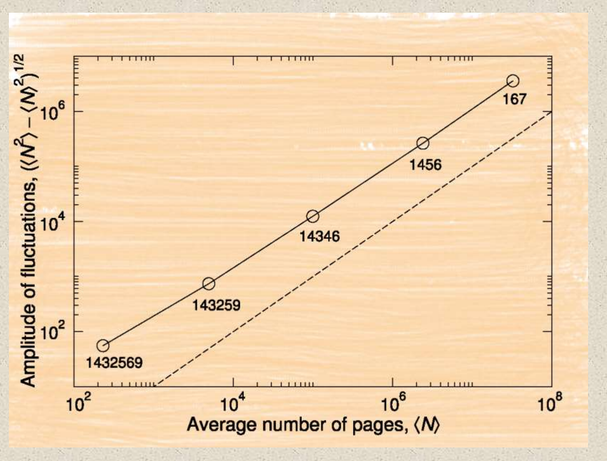

Figure 5: Amplitude of the fluctuations, $(\langle N^2\rangle - \langle N\rangle^2)^{1/2}$, of the frequencies of web pages holding non-round numbers vs. their mean values, $\langle N\rangle$, on a log-log plot. The data (circles) were obtained resorting to the same intervals as in Fig. 4. Next to each circle the average (non-round) number, $\langle n\rangle$, for the corresponding interval is indicated. The dashed line has slope 1, far from the ½ exponent observed in standard statistical fluctuations [3,4]. One can see that $(\langle N^2\rangle - \langle N\rangle^2)^{1/2} \approx 0.1\langle N\rangle$ for $\langle N\rangle > 10^3$.

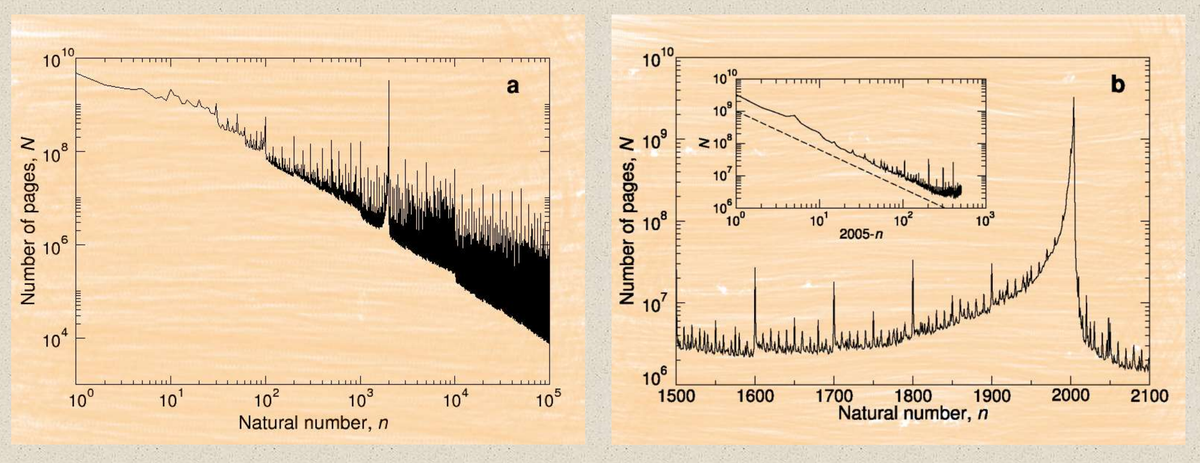

Figure 2: **a,** The amount N of web pages containing a number n. Note the peak in 2004 (the current year at the time the measurements were made). **b,** Blow up of the region of numbers associated to years. In the inset, the power-law decay as we move to the past.

These observations suggest a new view of the array of integers in the WWW (and in Nature) as a complex, evolving, inhomogeneous system. The statistics of numbers turns out to be far more rich and complex than one might expect based on classical Benford's law [2].

The global array of numbers is surmised to be a "numeric snapshot of the collective consciousness" [5]. So, the study of their statistics could lead to a better understanding of a wide circle of sociological and psychological phenomena. The distribution of numbers in human documents contains a wealth of diverse information in an integrated form. The detailed analysis of the general statistics of numbers in the WWW could allow the effective extraction and evaluation of this hidden information.

More details on this work can be found in [6].

References:

[1] S. Newcomb, *Amer. J. Math.* **4** (1881) 39
[2] F. Benford, *Proc. Amer. Phil. Soc.* **78** (1938) 551
[3] L. D. Landau & E. M. Lifshitz, *Statistical Physics, Part 1* (Pergamon Press, New York, 1993)
[4] M. Argollo de Menezes & A.-L. Barabási, *Phys. Rev. Lett.* **92** (2004) 028701
[5] G. Levin *et al. The secret lives of numbers* <http://www.turbulence.org/Works/nums/> (2002)
[6] S. N. Dorogovtsev, J. F. F. Mendes & J. G. Oliveira, physics/0504185 (to be published in Physica A)

Acknowledgements:

This work was partially supported by projects POCTI/FAT/46241/2002 and POCTI/MAT/46176/2002. S.N.D. and J.F.F.M. acknowledge the NATO program OUTREACH for support. J.G.O. acknowledges financial support of FCT (Portugal), grant No. SFRH/BD/14168/2003.