Frequency of occurrence of numbers in the World Wide Web

Sergey N. Dorogovtsev
Departamento de Física da Universidade de Aveiro, 3810-193 Aveiro, Portugal
A. F. Ioffe Physico-Technical Institute, 194021 St. Petersburg, Russia

José Fernando F. Mendes
Departamento de Física da Universidade de Aveiro, 3810-193 Aveiro, Portugal

João Gama Oliveira
Departamento de Física da Universidade de Aveiro, 3810-193 Aveiro, Portugal
Department of Physics and Center for Complex Network Research, University of Notre Dame, IN 46556, USA

Presentation: Poster
Topic: WWW, numbers, statistics

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.