Benford's Law in Social Networks

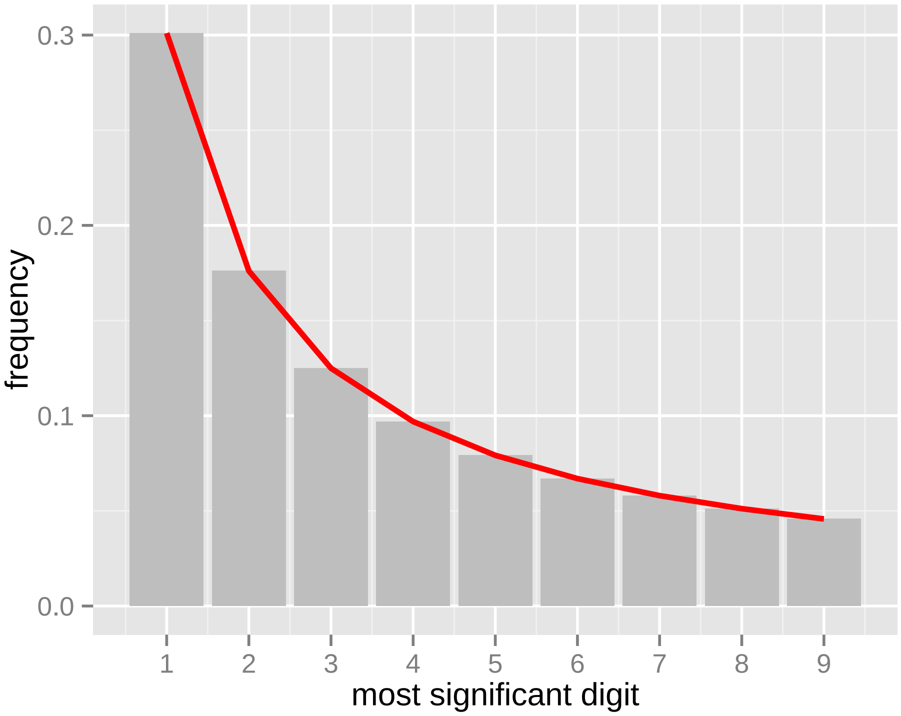
Mikołaj Morzy

December 14, 2015

Benford's Law is a well-documented phenomenon describing the distribution of the first digit in a large collection of numbers. Originally noticed by Newcomb and Benford, it estimates the probability of the most significant digit (msd) being d as

$$P(msd = d) = \log_{10}(1 + \frac{1}{d})$$

The graphical representation of Benford's Law is presented in the following Figure.



Benford's distribution of the leading digit has been observed in a very large number of datasets, both natural and synthetic. For instance, the lack of concordance with the Benford's distribution in accounting data is a strong indicator of possible fraud. Benford distribution has been observed in genomic data, in the geographical data (lengths of rivers, areas of lakes), but also in the set of most popular physical constants, in the observed and predicted 477 radioactive half-lives, in many population related datasets, in failure rates and mean-time-to-failure values in information systems, and many more. Benford's distribution is scale-invariant (i.e. the set of considered values may be multiplied by any constant and the distribution of most significant digits will still fit the Benford distribution), but it is also base-invariant (i.e. it is independent of the base of the counting system).

Recently it has been suggested [1] that Benford's Law applies to social networks, and in particular, to the distribution of vertex degrees. The author analyzes five datasets (Facebook, Twitter, Google+, Pinterest, LiveJournal) and concludes that the Benford distribution can be fitted to the distribution of vertex degrees in these networks. The concordance between the observed distribution of vertex degrees and the theoretical Benford distribution is decided based on the value of the Pearson's correlation coefficient - if it exceeds 0.97 the author assumed the presence of Benford's Law. Unfortunately, this approach is simply wrong. As a counter-example consider the following set of functions and the results of Pearson's correlation test for these functions (after normalization) and the Benford's distribution, along with the Kolmogorov-Smirnov goodness of fit p-values.

function	$f_1(x) = \frac{1}{x}$	$f_2(x) = x$	$f_3(x) = \frac{1}{2^x}$	$f_4(x) = (\frac{3}{4})^x$	$f_5(x) = \frac{1}{x+1}$
Pearson correlation	0.9965739	-0.8638019	0.9962061	0.9688657	0.9964537
KS test	0.9894694	0.9894694	0.1258741	0.9894694	0.9894694

For such small sample size (only 9 points representing each possible leading digit) almost every exponential function will be highly correlated with Benford's distribution. For small sample sizes the Kolmogorov-Smirnov test has very low statistical power, i.e. the fact that the test fails to reject the null hypothesis is non-informative due to very high number of false negatives. In addition, we can easily find distributions that have high Pearson correlation coefficient, high KS test p-value, or both, rendering the comparison meaningless.

The literature provides several methods for testing if a given set of values is a Benford Set, i.e. if the distribution of the most significant digits in the set follows the Benford distribution. Among the most common approaches are:

- performing the Pearson χ^2 test of the goodness of fit of categorical distributions
- mantissa arc test
- mean absolute deviation measure
- distortion factor

We have downloaded 10 different real world networks from the SNAP dataset [2] and we have computed the distributions of the following centrality measures:

- degree
- betweenness
- local clustering coefficient
- closeness

The datasets considered in our experiment were:

	network	#vertices	#edges
1	amazon	262111	1 234 877
2	citations	27 770	352 807
3	dblp	317 080	1 049 866
4	enron	36 692	367 662
5	facebook	4 039	88 234
6	gnutella	36 682	88 328
7	physics	12 008	237 010
8	slashdot	82 168	948 464
9	twitter	81 306	2 420 766
10	wikipedia	7 115	103 689

Then, for each of these centrality measures we have performed five independent tests: the Pearson χ^2 test of the goodness of fit, the mantissa arc test, the MAD, and the traditional Pearson's correlation and the distortion factor. We have assumed the significance level of $\alpha = 5\%$ where the results of the test are expressed as p-values. For mean absolute deviation we have followed the suggestion in [3] and we have concluded the presence of Benford distribution for $MAD < 0.0018$, and for the distortion factor we have compared the test statistic against the 95th percentile of the standard normal distritubion. Thus, for each network and for

	network	measure	#flags
1	dblp.csv	betweenness	3
2	enron.csv	betweenness	2
3	facebook.csv	betweenness	3
4	physics.csv	betweenness	3
5	slashdot.csv	betweenness	3
6	twitter.csv	betweenness	3
7	wikipedia.csv	betweenness	3

each centrality measure we have computed five flags signaling the presence of Benford distribution. We have decided to conclude that a given measure and a given network follows Benford's Law if any two of these flags were set. Below we list the networks and their measures which have passed our test.

As can be easily seen, the only measure of online social networks which is concordant with the Benford's Law is the betweenness. Two networks pass a signle test on the degree measure, but in both cases this is an unreliable Pearson's correlation coefficient (the same situation happened for the closeness measure). None of the observed distributions of the local clustering coefficients have passed any of the five tests.

We have also decided to verify if popular artificial network models produce Benford's distribution of most significant digits in distributions of centrality measures. We have selected the following models:

- random graph model
- small world model
- preferential attachment model
- forest fire model

For each model we have selected five different values of the main parameter. For the random graph model the main parameter is the random edge creation probability, for the small world model it is the rewiring probability, for preferential attachment model it is the exponent of the degree distribution, and for the forest fire model it is the forward burning probability. For each model and for each particular value of the main parameter we have generated 100 networks, each consisting of $n = 10\,000$ vertices. The values of all tests (we have performed the same tests as in the case of real world networks) were averaged over all 100 realizations. The results are presented below.

	model	parameter	measure
1	preferential.attachment	1.00	closeness
2	preferential.attachment	1.50	closeness
3	preferential.attachment	2.00	closeness
4	preferential.attachment	2.50	betweenness
5	preferential.attachment	2.50	closeness
6	preferential.attachment	3.00	betweenness
7	preferential.attachment	3.00	closeness
8	random.graph	0.01%	closeness
9	random.graph	0.0325%	closeness

As we can see, out of 80 possible combinations (4 models, 4 measures, 5 values of parameters) only 9 combinations pass the Benford's distribution concordance test (always passing the Pearson's χ^2 test and the mantissa arc test), and in the majority of combinations it is the closeness measure which follows the Benford's Law. Betweenness starts to follow the Benford's distribution only for the preferential attachment model, and only for large values of the degree distribution exponent.

Our conclusions are twofold:

- online social networks do not exhibit Benford's distribution in their centrality measures, with the exception of betweenness

- generative network models do not produce Benford's distribution in centrality measures except for closeness

[1] Golbeck J (2015) Benford's Law Applies to Online Social Networks. PLoS ONE 10(8): e0135169. doi:10.1371/journal.pone.0135169

[2] http://snap.stanford.edu/data/

[3] Nigrini M (2011) Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations, John Wiley & Sons