The Effect of Recency to Human Mobility – Supporting Information –

Hugo Barbosa[1,*], Fernando B. de Lima-Neto[2], Alexandre Evsukoff[3], and Ronaldo Menezes[1]

[1]BioComplex Lab, Department of Computer Sciences, Florida Institute of Technology, USA
[2]Computational Intelligence Research Group, Polytechnic School, University of Pernambuco, Brazil
[3]COPPE, Federal University of Rio de Janeiro, Brazil
[*]hbarbosa@biocomplexlab.org

A Statistical analysis of the rank distributions

As presented in the main text, the probability distributions of the rank variables K_f and K_s are very similar. In order to assess whether the two rank variables come from the same distribution, we performed the two-sided Kolmogorov-Smirnov (KS) test. To test our hypothesis (i.e., both variables come from the same distribution), we compare the Kolmogorov-Smirnov distance D_{K_f,K_s} with the critical value D_α for a desired α level where

$$D_\alpha = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}},$$

with sample size $n_1 = n_2 = n$ and $c(\alpha) = 1.95$ for $\alpha = 0.001$ [4, 5]. As one can see from the Table A1, the distance $D_{K_f,K_s} > D_\alpha$ for both datasets, rejecting the hypothesis that K_f and K_s were drawn from the same distribution.

Table A1: Two-sided Kolmogorov-Smirnov test and p-values for the rank variables.

Dataset	n	D_{K_f,K_s}	D_α	p-value
$D1$	564,228	0.07999	0.00367	0.0
$D2$	2,267,116	0.09708	0.00183	0.0

Additionally, in order to determine the probability function that better characterizes the rank variables, we compared the goodness of fit offered by different heavy-tailed distributions. For this test, we measured the fit provided by two other distributions, namely log-normal and double-Pareto log-normal (dPlN). In some scenarios, the log-normal and the truncated power law distributions can both yield very similar results.

More recently, the dPlN distribution has been reported to offer a very sound model for many empirical data such as income distribution, oil-field sizes [3] and the degree distribution of social networks [2]. The double-Pareto log-normal corresponds to a mixture of two power laws joined by a log-normal segment [3]. The PDF of the dPlN can be defined as

$$f(x) = \frac{\alpha\beta}{\alpha+\beta}\left[f_1(x) + f_2(x)\right], \tag{1}$$

$$f_1(x) = x^{-\alpha-1} e^{\alpha\nu+\alpha^2\tau^2/2}\Phi\left(\frac{\ln x - \nu - \alpha\tau^2}{\tau}\right),$$

$$f_2(x) = x^{\beta-1} e^{-\beta\nu+\beta^2\tau^2/2}\Phi^c\left(\frac{\ln x - \nu + \beta\tau^2}{\tau}\right),$$

where Φ is the CDF (Cumulative Distribution Function) of the standard normal $N(0,1)$ and Φ^c is the CCDF of $N(0,1)$.

The log-likelihood ratio test compares two competing candidate distributions where the one with the higher likelihood is the one that provides the better fit. The sign of the log-likelihood ratio indicates the prevalence of the target distribution (here, the truncated power law) over an alternative competing hypotheses whereas the p-value indicates the significance level of the test [1]. As shown on the Table A2, both rank variables are indeed better approximated by truncated power laws (see also Figure A1).

B Parameters estimation

The probability distribution of the rank-based variables described in the main text were better approximated by truncated power-law distributions $p(x) = Cx^{-\alpha}\mathrm{e}^{-x/\tau}$. Parameters were estimated using the methods in Ref. [1]. Table B3 shows the best parameters of the truncated power laws estimated by MLE.

C Randomized datasets

In order to verify whether the power law observed in the recency rank distribution is rooted on the temporal semantics of individuals' trajectories, we applied our rank-based approach to randomized versions of both empirical datasets ($D1$ and $D2$). The first randomized dataset we analyzed ($R1$) was obtained from uniformly shuffling each individual trajectory. This way, we artificially remove any temporal information possibly encoded within the individual trajectories,

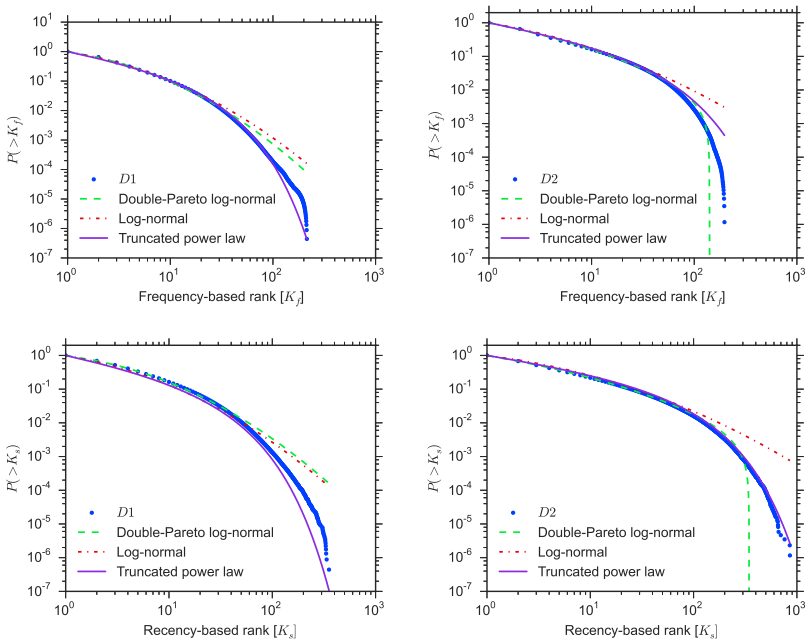

Figure A1: Curve fits of different heavy-tailed distributions for both K_f (top charts) and K_s (bottom charts). In addition to the well-known log-normal (dot-dashed line) and truncated power law (solid line) distributions, we also measured the goodness of fit for the double-Pareto log-normal.

while maintaining the visitation frequencies intact. On the second randomization method ($R2$), we also remove the visitation frequencies by generating for each user a new random trajectory with the same number of displacements, and the same number of distinct visited locations. To serve as the baseline for the analyses, the data of the third randomization approach ($R3$) produces a new dataset with the same size as the original one, but keeping only the total number of users and locations. More precisely, for each of the datasets, we generated a randomized version of them with M random points

$$v_m = [u_m, l_m, m], m \in [1, \ldots, M],$$

where each u_m, l_m is uniformly sampled from U users and N locations respectively, with M, U and L the same as in $D1$ and $D2$.

Table A2: Comparison between the goodness of fit provided by the truncated power law and other distributions via log-likelihood ratio test.

Dataset	Rank	Alternative distribution	Log-likelihood ratio	p-value
$D1$	frequency	Log-normal	50.195	0.0
		Double-Pareto Log-normal	157.185	0.0
		Exponential	227.77	0.0
		Stretched Exponential	29.40	0.0
	recency	Log-normal	46.147	0.0
		Double-Pareto log-normal	68.077	0.0
		Exponential	189.95	0.0
		Stretched Exponential	29.30	0.0
$D2$	frequency	Log-normal	114.455	0.0
		Double-Pareto log-normal	90.703	0.0
		Exponential	223.56	0.0
		Stretched Exponential	98.77	0.0
	recency	Log-normal	45.58	0.0
		Double-Pareto log-normal	128.884	0.0
		Exponential	218.98	0.0
		Stretched Exponential	30.029	0.0

Table B3: Estimated parameters of the truncated power-law distributions with the best fit for the rank variables.

Dataset	Rank	α	τ
$D1$	recency	1.644	41.66
	frequency	1.859	37.0
$D2$	recency	1.699	250.0
	frequency	1.625	125.0

References

[1] Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, November 2009.

[2] Seshadri Mukund, Sridhar Machiraju, Ashwin Sridharan, Jean Bolot, Christos Faloutsos, and Jure Leskovec. Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions. 2008.

[3] William J. Reed and Murray Jorgensen. The Double Pareto-Lognormal DistributionA New Parametric Model for Size Distributions. *Communications in Statistics - Theory and Methods*, 33(8):1733–1753, 2004.

[4] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Statist.*, 19(2):279–281, 06 1948.

[5] Nikolai V Smirnov. On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou*, 2(2), 1939.