P o C S What's The Story?

Principles of Complex Systems, Vols. 1 & 2, CSYS/MATH 300 and 303
University of Vermont, Fall 2021
Assignment 11
code name: Romeo v Juliet: Dawn of Justness

Due: Friday, November 19, by 11:59 pm, 2021.
Relevant clips, episodes, and slides are listed on the assignment's page:
https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems//assignments/11/
Some useful reminders:
Deliverator: Prof. Peter Sheridan Dodds (contact through Teams)
Assistant Deliverator: Michael Arnold (contact through Teams)
Office: The Ether
Office hours: TBD
Course website:
https://pdodds.w3.uvm.edu//teaching/courses/2021-2022principles-of-complex-systems

---

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you collaborated.

For coding, we recommend you improve your skills with Python, R, and/or Julia. The Deliverator uses Matlab.

Graduate students are requested to use LaTeX (or related TeX variant). If you are new to LaTeX, please endeavor to submit at least n questions per assignment in LaTeX, where n is the assignment number.

Assignment submission: Via Blackboard.

Please submit your project's current draft in pdf format via Blackboard by the same time specified for this assignment. For teams, please list all team member names clearly at the start.

---

1. (3 + 3 + 3)

   Estimating the rare:

   Google's raw data is for word frequency $k \geq 200$ so let's deal with that issue now.

   From Assignment 2, we had for word frequency in the range $200 \leq k \leq 10^7$, a fit for the CCDF of

   $$N_{\geq k} \sim 3.46 \times 10^8 k^{-0.661},$$

   ignoring errors.

(a) Using the above fit, create a complete hypothetical N_k by expanding N_k back for $k = 1$ to $k = 199$, and plot the result in double-log space (meaning log-log space).

(b) Compute the mean and variance of this reconstructed distribution.

(c) Estimate:

i. The hypothetical fraction of words that appear once out of all words (think of words as organisms or tokens here),

ii. The hypothetical total number and fraction of unique words in Google's data set (think at the species or type level now),

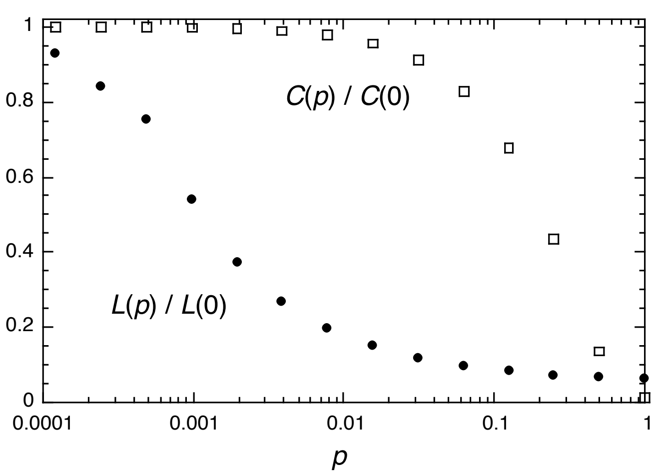
iii. And what fraction of total words are left out of the Google data set by providing only those with counts $k \geq 200$ (back to words as organisms or tokens).

2. Simulate the small-world model and reproduce Fig. 2 from the 1998 Watts-Strogatz paper showing how clustering and average shortest path behave with rewiring probability p [1].

   Please find and use any suitable code online, and feel free to share with each other via Slack.

   Use $N = 1000$ nodes and $k = 10$ for average degree, and vary p from 0.0001 to 1, evenly spaced on a logarithmic scale (there are only 14 values used in the paper).

   Here's the figure you're aiming for:



References

[1] D. J. Watts and S. J. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998. pdf