P o C S  What's The Story?

Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2014
Assignment 11 • code name: Don't know where, don't know when.

Dispersed: Friday, November 14, 2014.
Due: By start of lecture, 1:00 pm, Never, 2014.
Some useful reminders:
Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 2:30 pm to 3:45 pm on Tuesday, 12:30 pm to 2:00 pm on Wednesday
Course website: http://www.uvm.edu/∼pdodds/teaching/courses/2014-08UVM-300

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.

Graduate students are requested to use LaTeX (or related TeX variant).

---

1. Yes, even more on power law size distributions. It's good for you.

   For the probability distribution $P(x) = cx^{-\gamma}$, $0 < a \leq x \leq b$, compute the mean absolute displacement (MAD), which is given by $\langle |X - \langle X \rangle| \rangle$ where $\langle \cdot \rangle$ represents expected value. As always, simplify your expression as much as possible.

   MAD is a more reasonable estimate for the width of a distribution, but we like variance σ^2 because the calculations are much prettier. Really.

2. In the limit of $b \to \infty$, show that MAD asymptotically behave as:

$$\langle |X - \langle X \rangle| \rangle = \frac{2(\gamma - 2)^{(\gamma-3)}}{(\gamma - 1)^{(\gamma-2)}} a.$$

   How does this compare with the behavior of the variance? (See the last question of Assignment 1.)

3. *"Any good idea can be stated in fifty words or less."—Stanisław Ulam.*[1]

   The top of the narrative hierarchy:

   Read through Anderson's seminal paper "More is different" [1] and generate three descriptions of complexification with exactly the following lengths:

   (a) Three words,

[1] At the very least, Ulam's claim is self-consistent.

(b) Six words,

(c) and Twelve words.

Things have sped up since Ulam made his claim. All three may contain one or more sentences.

4. The next two questions continue on with the Google data set we first examined in Assignment 1.

   Using the CCDF and standard linear regression, measure the exponent $\gamma - 1$ as a function of the upper limit of the scaling window, with a fixed lower limit of $k_{\min} = 200$.

   Please plot γ as a function of $k_{\max}$, including 95% confidence intervals.

   Note that the break in scaling should mess things up but we're interested here in how stable the estimate of γ is up until the break point.

   Comment on the stability of γ over variable window sizes.

   Pro Tip: your upper limit values should be distributed evenly in log space.

5. (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 here),

      ii. the hypothetical total number and fraction of unique words in Google's data set (think at the species 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).

References

[1] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972. pdf