Principles of Complex Systems, CSYS/MATH 300—Assignment 1
University of Vermont, Fall 2010

Dispersed: Friday, September 17, 2010.
Due: By start of lecture, 1:00 pm, Thursday, September 24, 2010.
Some useful reminders:
Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 1:00 pm to 4:00 pm, Wednesday
Course website: `http://www.uvm.edu/~pdodds/teaching/courses/2010-08UVM-300`

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

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

All about power law distributions (basic computations and some real life data from the Big Googster):

1. Consider a random variable X with a probability distribution given by

   $$P(x) = cx^{-\gamma}$$

   where c is a normalization constant, and $0 < a \leq x \leq b$. (a and b are the lower and upper cutoffs respectively.) Assume that $\gamma > 1$.

   (a) Determine c.

   (b) Why did we assume $\gamma > 1$?

   Note: For all answers you obtain for the questions below, please replace c by the expression you find here, and simplify expressions as much as possible.

2. Compute the nth moment of X.

3. In the limit $b \to \infty$, how does the nth moment behave as a function of γ?

4. For finite cutoffs a and b with $a \ll b$, which cutoff dominates the expression for the nth moment as a function of γ and n?

   Note: both cutoffs may be involved to some degree.

5. (a) Noting what constraints, if any, we must place on γ for the mean to be finite in the case $b \to \infty$, find σ, the standard deviation of X.

(b) How does σ behave as a function of γ?

6. 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.

7. How does MAD behave as a function of γ? How does this compare with the variance?

8. Drawing on a Google vocabulary data set (see below for links),

   (a) Plot the frequency distribution N_k representing how many distinct words appear k times in this particular corpus as a function of k.

   (b) Repeat the same plot in log-log space (using base 10, i.e., plot $\log_{10} N_k$ as a function of $\log_{10} k$).

   (c) Using your eyeballs, indicate over what range power-law scaling appears to hold and, estimate, using least squares regression over this range, the exponent in the fit $N_k \sim k^{-\gamma}$.

   You are encouraged but not required to use matlab.

   The data (links are clickable):

   - Matlab file (`wordfreqs` $= k$, `counts` $= N_k$): http://www.uvm.edu/~pdodds/teaching/courses/2010-08UVM-300/docs/google_vocab_freqs.mat
   - Compressed text file (first column $= k$, second column $= N_k$): http://www.uvm.edu/~pdodds/teaching/courses/2010-08UVM-300/docs/vocab_cs_mod.txt.gz
   - Uncompressed text file (first column $= k$, second column $= N_k$): http://www.uvm.edu/~pdodds/teaching/courses/2010-08UVM-300/docs/vocab_cs_mod.txt

   Note: 'words' here include any separate textual object including numbers, websites, html markup, etc.

   Note: To keep the file to a reasonable size, the minimum number of appearances is $k_{\min} = 200$ corresponding to $N_{200} = 48030$ unique words.