Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2011
Assignment 2

Dispersed: Tuesday, September 20, 2011.
Due: By start of lecture, 11:30 am, Thursday, September 29, 2011.
Some useful reminders:
Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 12:50 pm to 3:50 pm, Wednesday
Course website: http://www.uvm.edu/∼pdodds/teaching/courses/2011-08UVM-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).

For Q1–5, you'll further explore the Google data set you examined for the first problem set. For Q6–7, you'll examine how the largest sample size grows with the number of samples.

1. Plot the complementary cumulative distribution function (CCDF).
2. Using standard linear regression, measure the exponent $\gamma - 1$ where γ is the exponent of the underlying distribution function. Identify and use a range of frequencies for which scaling appears consistent. Report the 95% confidence interval for your estimate.
3. Plot word frequency as a function of rank in the manner of Zipf.
4. Using standard linear regression, measure α, Zipf's exponent. Report the 95% confidence interval for your estimate.
5. Write down how γ and α are related and check how this expression works for your estimates here.
6. (3+3)

   Consider a set of N samples, randomly chosen according to the probability distribution $P_k = ck^{-\gamma}$ where $k \geq 1$ and $2 < \gamma < 3$. (Note that k is discrete rather than continuous.)

(a) Estimate $\min k_{\text{max}}$, the approximate minimum of the largest sample in the network, finding how it depends on N.

(Hint: we expect on the order of 1 of the N samples to have a value of $\min k_{\text{max}}$ or greater.)

(b) Determine the average value of samples with value $k \geq \min k_{\text{max}}$ to find how the expected value of k_{max} (i.e., $\langle k_{\text{max}} \rangle$) scales with N.

7. Let's see how well your answer for the previous question works. For $\gamma = 5/2$, generate $n = 100$ sets each of $N = 10$ samples, $N = 100$ samples, $N = 1000$ samples, $N = 10^4$ samples, and further if possible using $P_k = ck^{-5/2}$.

   For each set, find the maximum value. Then find the average maximum value for each N for each of the $n = 100$ instances. Plot $\langle k_{\text{max}} \rangle$ as a function of N and calculate the scaling.

   Key question: how do we computationally sample from a discrete probability distribution?