Principles of Complex Systems, CSYS/MATH 300—Assignment 5
University of Vermont, Fall 2010

Dispersed: Monday, October 25, 2010.
Due: By start of lecture, 1:00 pm, Thursday, November 4, 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).

---

1. In Simon's original model, the expected total number of distinct groups at time t is ρt. Recall that each group is made up of elements of a particular flavor.

   In class, we derived the fraction of groups containing only 1 element, finding

$$n_1^{(g)} = \frac{N_1(t)}{\rho t} = \frac{1}{2-\rho}$$

   (a) Find the form of $n_2^{(g)}$ and $n_3^{(g)}$, the fraction of groups that are of size 2 and size 3.

   (b) Using data for James Joyce's Ulysses (see below), first show that Simon's estimate for the innovation rate $\rho_{\text{est}} \simeq 0.115$ is reasonably accurate for the version of the text's word counts given below. You should find $\rho_{\text{est}} \simeq 0.119$

   (c) Now compare your theoretical estimates for $n_1^{(g)}$, $n_3^{(g)}$, and $n_3^{(g)}$, with empirical values you obtain for Ulysses.

   The data (links are clickable):

   - Matlab file (`sortedcounts` = word frequency f in descending order, `sortedwords` = ranked words):
     http://www.uvm.edu/∼pdodds/teaching/courses/2010-08UVM-300/docs/ulysses.mat
   - Colon-separated text file (first column = word count f, second column = word): http://www.uvm.edu/∼pdodds/teaching/courses/2010-08UVM-300/docs/ulysses.txt

Data taken from http://www.doc.ic.ac.uk/ rac101/concord/texts/ulysses/. Note that some matching words with differing capitalization are recorded as separate words.

2. *Zipfarama via Optimization:*

   Complete the Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, \ldots, p_n) = F(p_1, p_2, \ldots, p_n) + \lambda G(p_1, p_2, \ldots, p_n)$$

   where the 'cost over information' function is

$$F(p_1, p_2, \ldots, p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i \ln(i+a)}{-g \sum_{i=1}^{n} p_i \ln p_i}$$

   and the constraint function is

$$G(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i - 1 = 0$$

   to find

$$p_j = (j+a)^{-\alpha}$$

   where $\alpha = H/gC$.

   Note: We have now allowed the cost factor to be $(j+a)$ rather than $(j+1)$. Exciting!

   Hint: when finding λ, find an expression connecting λ, g, C, and H. Extra hint: one way may be to substitute the form you find for $\ln p_i$ into H's definition (but do not replace p_i).

3. (a) For $n \to \infty$, use some computation tool (e.g., Matlab, an abacus, but not a clever friend who's really into computers) to determine that $\alpha \simeq 1.73$ for $a = 1$. (Recall: we expect $\alpha < 1$.)

   (b) For finite n, find an approximate estimate of a in terms of n that yields $\alpha = 1$.

   (Hint: use an integral approximation for the relevant sum.)

   (c) What happens to a as $n \to \infty$?