P o C S What's The Story?

Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2013
Assignment 9 • code name: "Good evening Fräulein."

Dispersed: Thursday, November 14, 2013.
Due: By start of lecture, 1:00 pm, Thursday, November 21, 2013.
Some useful reminders:

Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu

Office hours: 10:30 am to 11:30 am, Monday, and 1:00 pm to 3:00 pm, Wednesday
Course website: http://www.uvm.edu/∼pdodds/teaching/courses/2013-08UVM-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).

Optional.

1. (3 + 3)

   Using Gleeson and Calahane's iterative equations below, derive the contagion condition for a vanishing seed by taking the limit $\phi_0 \to 0$ and $t \to \infty$. In lectures, we derived the discrete evolution equations for the fraction of infected nodes ϕ_t and the fraction of infected edges θ_t as follows:

$$\phi_{t+1} = \phi_0 + (1 - \phi_0) \sum_{k=0}^{\infty} P_k \sum_{j=0}^{k} \binom{k}{j} \theta_t^j (1 - \theta_t)^{k-j} B_{kj},$$

$$\theta_{t+1} = G(\theta_t; \phi_0) = \phi_0 + (1 - \phi_0) \sum_{k=1}^{\infty} \frac{k P_k}{\langle k \rangle} \sum_{j=0}^{k-1} \binom{k-1}{j} {\theta_t}^j (1 - \theta_t)^{k-1-j} B_{kj},$$

   where $\theta_0 = \phi_0$, and B_{kj} is the probability that a degree k node becomes active when j of its neighbors are active.

   Recall that by contagion condition, we mean the requirements of a random network for macroscopic spreading to occur.

   To connect the paper's model and notation to those of our lectures, given a specific response function F and a threshold model, the B_{kj} are given by $B_{kj} = F(j/k)$.

   Allow B_{k0} to be arbitrary (i.e., not necessarily 0 as for simple threshold functions).

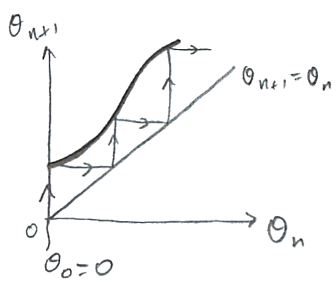
We really only need to understand how θ_t behaves. Write the corresponding equation as $\theta_{t+1} = G(\theta_t; \phi_0)$ and determine when

(a) $G(0; \phi_0) > 0$ (spreading is for free).

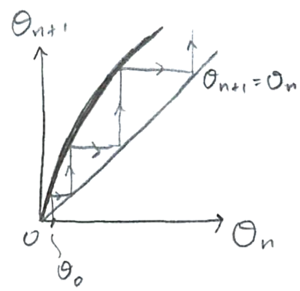
(b) $G(0; \phi_0) = 0$ and $G'(0; \phi_0) > 1$ meaning $\phi = 0$ is a unstable fixed point.

Here's a graphical hint for the three cases you need to consider as $\theta_0 \to 0$:

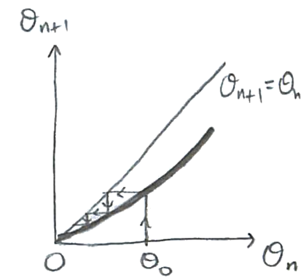
Success:



Sucesss:



Fail:



2. (3 + 3 + 3 + 3 + 3) *More on the power law stuff:*

Take x to be the wealth held by an individual in a population of n people, and the number of individuals with wealth between x and $x + \mathrm{d}x$ to be approximately $N(x)\mathrm{d}x$.

Given a power-law size frequency distribution $N(x) = cx^{-\gamma}$ where $x_{\min} \ll x \ll \infty$, determine the value of γ for which the so-called 80/20 rule holds.

In other words, find γ for which the bottom 4/5 of the population holds 1/5 of the overall wealth, and the top 1/5 holds the remaining 4/5.

Assume the mean is finite, i.e., $\gamma > 2$.

(a) Determine the total wealth W in the system given $\int_{x_{\min}}^{\infty} \mathrm{d}x N(x) = n$.

(b) Imagine that $100q$ percent of the population holds $100(1 - r)$ percent of the wealth.

Show γ depends on p and q as

$$\gamma = 1 + \frac{\ln \frac{1}{(1-q)}}{\ln \frac{1}{(1-q)} - \ln \frac{1}{r}}.$$

(c) Given the above, is every pairing of q and r possible?

(d) Find γ for the 80/20 requirement.

(e) For the 80/20 γ you find, determine how much wealth $100q$ percent of the population possesses as a function of q and plot the result.

3. 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.

4. (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).