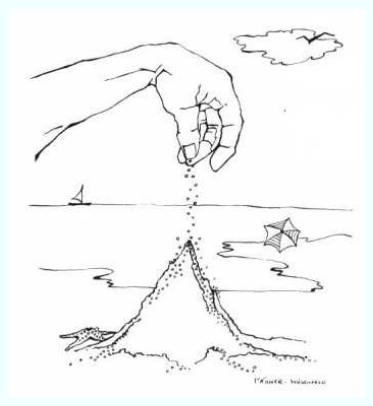
Inference, Models and Simulation for Complex Systems
CSCI 7000-003, Fall 2010
Prof. Aaron Clauset
Problem Set 2, due 9/29



1. **The sand pile model**

   One of the original models of self-organized criticality was the sand pile model, by Bak, Tang and Wiesenfeld in 1988. This model assumes a d-dimensional lattice where each cell has a capacity to hold c grains of sand. If at any time a cell has more than c grains, the stack of sand "topples" and the c grains are distributed evenly among that cell's neighbors, which may, in turn, cause them to topple. This toppling process repeats at every location on the lattice until all cells are under capacity. At each time step, exactly one grain of sand is added to the lattice and all dynamics are done before the next grain is added. Finally, there is at least one special "absorbing" cell on the lattice that destroys any sand grains that topple onto it. An *avalanche* of size x is defined as a time step during which x topple events occurred.

   (a) In the canonical sand pile model, the lattice is $n \times n$ and $c = 4$ so that when a stack topples, it distributes exactly one grain of sand to each of the north, east, south and west neighbors. New grains of sand are added only to the center-most cell of the lattice and any grains that topple off the edge of the lattice are destroyed (imagine the lattice as a table and sand that spills off the edge of the table is lost).

   Write a simulation of this canonical sand pile model. Choose n sufficiently large to demonstrate the self-organized critical behavior. Let the system converge to that state and then measure the distribution of avalanche sizes $\Pr(x)$. Plot your result as a ccdf on log-log axes. Label your axes.

(b) For small values of x, discrete power-law distributions are slightly different from the continuous version and as a result, the continuous MLE can yield inaccurate estimates of α when applied to discrete data.[1] The log-likelihood function for the discrete power law is

$$\mathcal{L}(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i \ , \tag{1}$$

where $\zeta(\alpha, x_{\min}) = \sum_{i=x_{\min}}^{\infty} i^{-\alpha}$ and is called the incomplete Zeta function.

Estimate α for your avalanche data by setting $x_{\min} = 1$ and numerically maximizing Eq. (1).

(c) (**optional**) Although the sand pile model is not a particularly realistic model of anything, its "toppling" dynamics and avalanches are a bit like the kind of cascading failures that can be seen in the electrical power grid. For example, if one transmission line becomes overloaded and fails, it sheds its load onto other nearby transmission lines, which themselves might become overloaded and fail, and so on, potentially leading to a massive system-wide collapse. Intuitively, having more unused capacity in the system will reduce the frequency of big cascades, but greater unused capacity also reduces the overall efficiency of the system.

Let the "efficiency" of the system be defined as the average fraction of the total sand capacity that is occupied in the critical state, that is,

$$E = \left\langle \frac{1}{c\, n^2} \sum_{i,j} x_{ij} \right\rangle_t \ , \tag{2}$$

where x_{ij} is the number of grains of sand on the i,jth cell, and we're averaging over a long period of time t. Measure and plot the efficiency of the sand pile as a function of the cumulative number of sand grains added. Label your axes.

Now think about how you could *increase* the efficiency of the sand pile model without changing its fundamental dynamics. Try one of your ideas. How does the efficiency change? How does the distribution of avalanche sizes change? Present and comment on your results.

[1] For large $x_{\min}$, this difference becomes small, and it's okay to approximate discrete data as continuous data. In your simulation, there's also a maximum avalanche size of roughly $s_{\max} = n^2$, which imposes a cutoff in the upper end of the distribution; we'll ignore this detail.

2. **The drunkard's walk**

Consider a very drunk person standing at the edge of a cliff with an infinite plane stretching out in the opposite direction. This person is so perfectly drunk that with equal probability they either take a single step toward the cliff (left) or away from the cliff (right). This is equivalent to a unbiased random walk on a 1d lattice (or equivalently, a simple random walk on the non-negative integers $0, 1, 2, \dots$).

(a) (**optional**) Derive the expected waiting time for the drunkard to fall off the cliff, that is, to return to 0.

(b) Set up and run a numerical experiment to estimate the expected lifetime of the drunkard. That is, simulate a very large number of drunkards, tabulate the distribution of their lifetimes, and then compute the average and its standard error.[2] Plot the distribution and report your estimate and its uncertainty. Label your axes. Comment on the differences between this process and a standard Poisson process. (Hint: think about hazard functions.)

(c) Now consider a "windy" variation. With small probability p, a gust of wind blows the drunkard k steps toward the cliff; otherwise, the drunkard takes a single step away from the cliff. Choose some p, simulate this process and tabulate the average lifetime of the drunkard as a function of k. Present and discuss your results.
(Hint: you'll want to try several values of p, since obviously, for large values, the gusts completely dominate the dynamics of the walk. Instead, we're interested in a more intermediate regime, with small-ish p, where the drunkard can make some progress in walking away from the cliff. **optional**: If you're feeling ambitious, either conduct a numerical experiment to characterize the impact of p on the dynamics, or do it analytically.)

(d) (**optional**) In the previous problem, the wind was blowing the drunkard toward the cliff. Now consider wind blowing *away* from the cliff, and the drunkard walking toward it. That is, consider a situation where occasionally a very large positive step is taken, but normally small negative steps are taken. This kind of model basically mixes a slow but strong growth process with a steady but weak decay process, and has been used to model the dynamics of viral populations under active counter-attack. Choose some p, simulate this process and characterize the average lifetime of the drunkard as a function of the size of the gust k. Present and discuss your results.

[2]The standard error for an estimated average $\langle x \rangle$ is defined as $\sigma / \sqrt{n}$ where n is the number of measurements in the average and σ is their standard deviation.

3. **Data analysis**

(a) Download one of data set 2A, 2B and 2C from the course webpage. Using the time-series analysis tools we discussed in class, determine the underlying structure of the generating random walk. Present your results, describe what you did to get there, and describe the generating process. Include a meaningful visualization of the empirical time series, and any figures showing the underlying structure you discovered.

(b) Simulate your hypothesis and compare a single instance of simulated data (i.e., a set of n synthetic observations $\{y_i\}$) to the empirical data. Present the comparison as a figure, and discuss the similarity and any discrepancies.
(Hint: run the same analytic treatment from part (a) on your simulated data to show that the simulated structure is the same as the empirical structure.)

(c) (**optional**) Construct and carry out a statistical hypothesis test to determine if your hypothesized generative process is a plausible explanation of the data, that is, compute a p-value for your hypothesis.
(Hint: remember that a good hypothesis test requires choosing a meaningful measure of deviations between the model and the data. Comparing $\{x_i\}$ and $\{y_i\}$ directly will probably not be a sufficiently severe measure of those deviations.)

(d) (**optional**) Download data set 2D from the course webpage. These data are a set of observed sequential pairs $\{(x_t, x_{t-1})\}$ for an unknown stochastic process; they are not ordered. Construct a size-dependent fluctuation scatterplot. Investigate its statistical structure. Present your results. (These particular data are pairs of estimated masses for extinct North American terrestrial mammals; x_t is the mass of a descendent species and x_{t-1} is the mass of that species' ancestor. They're not ordered because they're taken from a number of different lineages.)