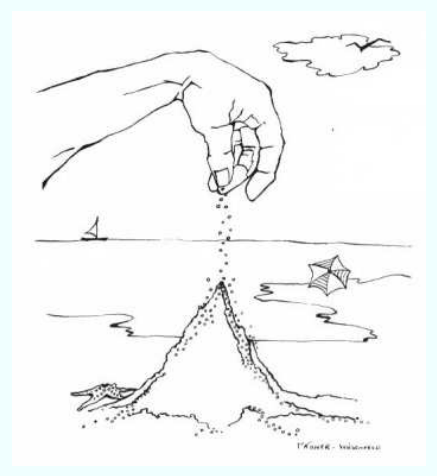
Inference, Models and Simulation for Complex Systems
CSCI 7000-001, Fall 2011
Prof. Aaron Clauset
Problem Set 2, due 9/27



1. **The Sciences of the Artificial** (15 points)

   Read Chapters 2 and 3. Write no more than two paragraphs that clearly and concisely explain why Simon thinks that simple models of human behavior are likely to be good models.

2. **The sand pile model** (40 points)

   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 grains are distributed to cell's neighbors—one grain of sand to each of them—which may, in turn, cause these neighboring cells to topple. This toppling process repeats at every location on the lattice until all cells are once again under capacity. At each time step, exactly one grain of sand is added to the lattice and all dynamics are completed 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. The self-organized critical behavior emerges once the table is "full" of sand; at this point, the distribution of avalanche sizes will follow a power law.

(a) (20 pts) 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. (If a cell has $k > c$ grains when it topples, it continues toppling until it returns to the stable $k < c$ state, e.g., if because of other toppling cells it accumulates $k = 9$ grains, it would topple twice, distributing 2 grains of sand to each of the four neighboring cells and have 1 grain remaining.) 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).[1]

Write a simulation of this canonical sand pile model. Choose n sufficiently large to demonstrate the self-organized critical behavior. Make three pretty figures showing the state of the system after $t = \{n, n^2, 4n^2\}$ grains of sand have been added. (Do you see why I chose these values?)

(b) (10 pts) Now, let the system converge to the self-organized critical state and measure the distribution of avalanche sizes $\Pr(x)$. Collect enough avalanches to get a satisfyingly stable empirical distribution.

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.[2] 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). Plot the empirical distribution of avalanche sizes (as a ccdf on log-log

[1] A simple variation of the canonical sand pile model chooses a cell uniformly at random on the lattice and adds a single grain there. Another variation uses "toroidal" boundary conditions—the left-edge is made to wrap around to touch the right-edge, and similarly the top and bottom edges—and makes the center-most cell the "absorbing" cell. Surprisingly, although this variation has many fewer absorbing cells than the canonical version, the avalanche dynamics are identical.

[2] 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.

axes) along with your fitted distribution. Report your value for $\hat{\alpha}$. No credit if you don't label your axes and provide a legend.

(c) (**10 pts extra credit**) For a spatial system like this, it is interesting to examining how correlated are the states (number of sand grains after all avalanches in a time step are complete) of two cells, at equilibrium, as a function of how distantly separated they are in the system. That is, if I tell you that cell A and B are a distance ℓ apart in the system and that A is in state s_A, the correlation function tells you how confident you can be that B is also in the same state.

Let the "distance" between two cells be their Manhattan distance, i.e., $d(M_{a,b}, M_{c,d}) = |a-c| + |b-c|$. If we let $N(\ell)$ denote the number of pairs of cells that are a distance ℓ apart, the correlation function is then defined as

$$\begin{aligned} C(\ell) &= \frac{E\left[(A-\mu_A)(B-\mu_B)\right]}{\sigma_A \sigma_B} \\ &= \frac{\sum_i \left(A_i - \langle A \rangle\right)\left(B_i - \langle B \rangle\right)}{\sqrt{\sum_i \left(A_i - \langle A \rangle\right)^2 \left(B_i - \langle B \rangle\right)^2}} \ , \end{aligned} \tag{2}$$

where A, B are elements such that $d(A, B) = \ell$. In words, we simply take all cells that are a distance ℓ from each other and we compute the (Pearson product-moment) correlation between their states.

Measure and plot the correlation function once your system has converged on the self-organized critical state. No credit if you don't label your axes and provide a legend. Write two sentences that explain why the function's form is interesting and how it reflects the self-organized state.

(d) (**10 pts extra credit**) 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,

$$E = \left\langle \frac{1}{c\,n^2} \sum_{i,j} x_{ij} \right\rangle_t \quad , \tag{3}$$

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.

3. **The drunkard's walk** (20 points)

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). We can model this process as an unbiased random walk on a 1-dimensional lattice, or equivalently as a random walk on the non-negative integers $0, 1, 2, \ldots$ in which $x_{t+1} = x_t + \lambda$ where $\Pr(\lambda = +1) = \Pr(\lambda = -1) = \frac{1}{2}$.

(a) (10 pts) 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.[3] Plot the distribution in a way that makes the distribution look straight, and report your estimate of the average and its uncertainty. Write 2-3 sentences about the differences between this process and a standard Poisson process, where an "event" is the death of the drunkard. (Hint: think about hazard functions.) No credit if you don't label your axes and provide a legend.

(b) (10 pts) 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. As a warm-up, choose a few interesting values of p, simulate this process and tabulate the average lifetime of the drunkard as a

[3]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.

function of k for each value.

Then, systematically analyze the dependence of the average lifetime on p (probability of a gust), k (strength of a gust) and T (length of the simulated walk). Make a single figure showing a smooth curve representing this dependence. Identify and discuss any obvious transitions between different dynamical regimes, and the behavior at extreme values. (Hint: notice that kp and $\langle t_{\text{cliff}} \rangle / T$ have special meanings.)

4. **Data analysis** (35 points)

   Here, we'll apply our time-series analysis techniques to studying a synthetic data set, and to one empirical data set. In all parts, no credit if you don't label your axes and provide a legend on your figures

   (a) (5 pts) Download one of PS2 data sets 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..
   (**5 pts extra credit** for each additional data set you analyze.)

   (b) (10 pts) 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.
   (**5 pts extra credit** for each additional data set you analyze.)
   (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) (10 pts) 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.
   (**5 pts extra credit** for each additional data set you analyze.)
   (Hint: remember that a good hypothesis test requires choosing a meaningful measure of deviations between the model and the data. Directly comparing the raw time series $\{x_i\}$ and $\{y_i\}$ will probably not be a sufficiently severe measure of those deviations.)

   (d) (10 pts) Download data set 2D from the course webpage. This data file gives the daily values of the Dow Jones Industrial Average stock index from 1 October

1928 to 15 September 2011 (downloaded from `finance.yahoo.com`). Extract and make a plot showing the time series of closing values. Use the tools from class to investigate its statistical structure. Present your results and explain what you find. (Hint: it may be useful to take logarithms.)

(**15 pts extra credit** if you can identify an underlying time-series process that yields a non-significant p-value under a hypothesis test, i.e., a process that is a plausible statistical explanation for the observed structure.)