

Due October 29 2009 in class

CS322: Network Analysis

Problem Set 2 - Fall 2009

If you have any questions regarding the problems set, send an email to the course assistants: simlac@stanford.edu and peleato@stanford.edu. Please write the name of your collaborators on your problem set. You can use existing software or code to compute the answers, you don't have to submit the source code.

The Problems

Problem 2.1

(From Easley and Kleinberg, Networks) In the basic “six degrees of separation” question, one asks whether most pairs of people in the world are connected by a path of at most six edges in the social network, where an edge joins any two people who know each other on a first-name basis.

Now let's consider a variation on this question. Suppose that we consider the full population of the world, and suppose that from each person in the world we create a directed edge only to their ten closest friends (but not to anyone else they know on a first-name basis). In the resulting “closest-friend” version of the social network, is it possible that for each pair of people in the world, there is a path of at most six edges connecting this pair of people? Explain.

Solution: In the described network, there will be a pair of people such that there is no path of at most six edges connecting them. Let us fix a person, p , in the network and consider the set of people who are within 6 steps from that person. The largest size of this set will occur in the case of a tree rooted at that person. So, the largest size (assuming directed edges) is the following;

$1(\text{person } p) + 10$ (num. of people in distance 1) $+ 100$ (num. of people in distance 2) $+ 1000 + 10000 + 100000 + 1000000 = 1111111$, which is clearly lot less than the world population (6 billion).

Hence, such a graph cannot connect every two people by a path of at most 6 edges. ■

Problem 2.2

You are developing a protocol to establish a peer-to-peer overlay network among n nodes. This protocol operates as follows.

Step 1: Each node flips a coin $(n-1)$ times to decide whether it generates an edge to each of the other $(n-1)$ nodes. The probability of doing so is p . Links are assumed undirected, regardless of which side establishes them. If two nodes flip their corresponding coins and both decide to connect to each other, only one edge is created.

Step 2: After this is done, every node not yet connected selects another node at random and establishes a link to this node.

If you let $p = \log n / (2n)$, does this protocol establish a connected network for large n ? (Hint: determine what small components exist after Step 1, and in particular, the number of isolated vertices.)

What would your answer be if p was only $1/n$?

Solution:

[We had originally thought of a different solution, but Stephen Dean Guo came up with the idea for the better one below]

If each side can establish an edge with probability p , the probability of any given edge existing in the network is $2p - p^2$. We realize that $2\frac{\log(n)}{2n} - \left(\frac{\log(n)}{2n}\right)^2 \rightarrow \frac{\log(n)}{n}$ when n tends to infinity, so we can assume that our graph is a $G(n, \frac{\log(n)}{n})$, i.e., the probability of any edge being present is $\frac{\log(n)}{n}$ (henceforth we will call this p). You might remember that this is exactly the threshold for connectivity of a random graph, so the proof will be somehow trickier than any other case. Some of you expressed concern over the theorem stating that $\forall \epsilon > 0$ the Erdos-Renyi graph with $p = (1 - \epsilon)\frac{\log(n)}{n}$ is disconnected. However, the p^2 term we neglected above cannot be viewed as that ϵ , since the ϵ is supposed to be a small CONSTANT greater than zero, and p^2 decreases with n .

Let k_m be the expected number of disconnected components of size m . Given a subset of m nodes, they will be disconnected from the rest iff all $m(n-m)$ edges between them and the rest of the graph are missing. The probability of this happening is $\left(1 - \frac{\log(n)}{n}\right)^{m(n-m)}$. On the other hand, the probability that all m nodes form a single component can be bounded using Cayley's theorem (The number of different spanning trees in a set of m nodes is m^{m-2}). Any connected component with m nodes will contain at least one spanning tree. Therefore we have the following chain of upper bounds:

$$\begin{aligned} P(m \text{ nodes are connected}) &\leq P(\text{there is a spanning tree}) \\ &\leq \sum_{i=1}^{m^{m-2}} P(\text{spanning tree number } i \text{ is present}) \\ &= m^{m-2} p^{m-1} \end{aligned}$$

where the second inequality comes from the union bound, and the last equality from the fact

that all spanning trees have the same number of edges ($m-1$).

Taking into account that there are $\binom{n}{m}$ possible subsets of m nodes we finally get,

$$k_m \leq u_m = \binom{n}{m} \left(1 - \frac{\log(n)}{n}\right)^{m(n-m)} m^{m-2} p^{m-1}.$$

We found an upper bound for k_m , which we will call u_m for reasons that will become clear later.

Massaging a bit the above expression and taking limits for large n , we get

$$\begin{aligned} k_m &\leq \frac{n^m}{m!} m^{m-2} e^{-\log(n)m \frac{n-m}{n}} \left(\frac{\log(n)}{n}\right)^{m-1} \\ &= \frac{m^{m-2}}{m!} n^{1-m} \log(n)^{m-1} \end{aligned}$$

Hence, for large n , $k_1 = 1$ and $k_m = 0$ for all $m > 1$. Step 2 will take care of the isolated node, and the expected number of larger components being isolated goes to zero. Unfortunately, this is not yet enough to assure that there will be no isolated components. Since the size of the possible components increases with n , we need to prove that their probability decreases fast enough so that $\sum_{i=2}^{\frac{n}{2}} k_i$ goes to zero. [For example, if we had $k_m = \frac{2}{n} \forall m$, then the expected number of isolated components of size m would be 0 for all m , but the expected number of isolated components of any size would be 1!!!]

We know that $\sum_{i=2}^{\frac{n}{2}} k_i \leq \sum_{i=2}^{\frac{n}{2}} u_i$. Lets find the ratio between u_{m+1} and u_m when n tends to infinity:

$$\begin{aligned} \frac{u_m}{u_{m+1}} &= \frac{\binom{n}{m} \left(1 - \frac{\log(n)}{n}\right)^{m(n-m)} m^{m-2} \left(\frac{\log(n)}{n}\right)^{m-1}}{\binom{n}{m+1} \left(1 - \frac{\log(n)}{n}\right)^{(m+1)(n-m-1)} (m+1)^{m-1} \left(\frac{\log(n)}{n}\right)^m} \\ &= \frac{(m+1)m^{m-2}}{(n-m)(m+1)^{m-1}} \left(1 - \frac{\log(n)}{n}\right)^{2m-n+1} \frac{n}{\log(n)} \\ &= \frac{(m+1)m^{m-2}}{(m+1)^{m-1}} \frac{n}{\log(n)} \end{aligned}$$

Thus, the expected number of isolated components of size m decreases as $\frac{\log(n)}{n}$ with each increment of m . Neglecting the constants, we can then bound the sum as:

$$\sum_{i=2}^{\frac{n}{2}} k_i \leq \sum_{i=2}^{\frac{n}{2}} u_i \leq k_2 \sum_{i=0}^{\frac{n}{2}} \left(\frac{\log(n)}{n}\right)^i < k_2 \sum_{i=0}^{\infty} \left(\frac{\log(n)}{n}\right)^i = k_2 \frac{1}{1 - \frac{\log(n)}{n}}$$

which tends to zero as n tends to infinity.

Finally, let's study the case of $p = \frac{2}{n}$. Given any two nodes, the probability that they are disconnected from the rest and connected to each other is $\frac{2}{n} \left(1 - \frac{2}{n}\right)^{2(n-2)}$ which is always larger than $\frac{2e^{-4}}{n}$. This probability tends to zero, but since the number of possible pairs increases with the number of nodes as $O(n^2)$, a constant fraction of the nodes will form isolated pairs (which step 2 will not reconnect).

■

Problem 2.3 Generate a dataset of 1 million values following a power-law distribution with exponent 2.5. Then compute experimentally the exponent of the distribution, using the following 4 methods:

Refer to *Power-law distributions in empirical data* by Clauset, Shalizi and Newman for how to generate random numbers from a power-law distribution.

- Fitting a line to the frequency distribution.
- Fitting a line to the frequency distribution with logarithmic binning.
- Using the complementary CDF.
- Using the maximum likelihood estimate.

Solution:

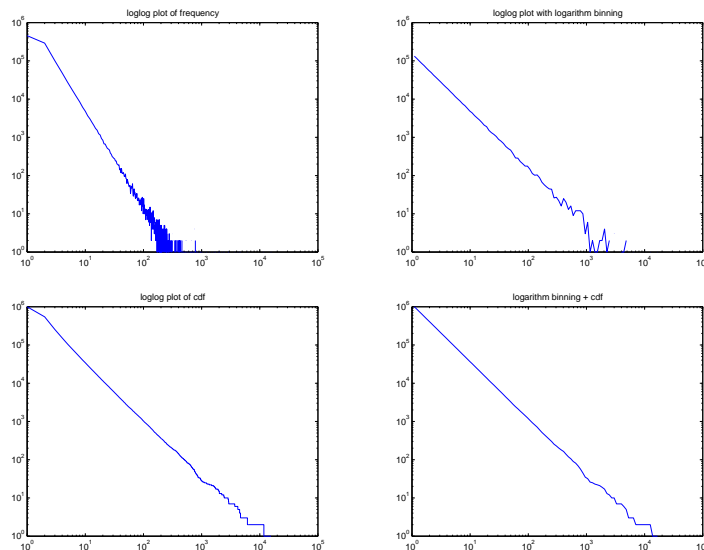


Figure 1: Plots for exponent estimation

The data is generated by generating a vector r of 10^6 numbers uniformly from $[0, 1]$ and apply the transformation $x = (1 - r)^{-2/3}$. We work with the continuous model in this problem. The calculation for discrete model is very similar. See Figure 1 for the plots.

- By setting bins of width 1 and doing linear regression of the frequencies in the loglog scale we get $\alpha = 0.2294$. The problem is that in the tail there are a lot of empty bins, so the

linear regression fits a flat line.

(b) Let bin i be $[1.1^{i-1}, 1.1^i]$. We count the frequency in each bin and normalize it by the width of the bin. Now by linear regression in the loglog scale we get $\alpha = 1.7895$. We obtained a total of 102 bins and the noise in the tail is not negligible. If we use only the first 60 bins for regression then the answer is very accurate ($\alpha = 2.5027$). Also it should be noted that if the counts for each bin is not normalized, we get a better estimate $\alpha = 2.3614$. This is one of the weird effect of those empty bins.

(c) Here we compute the CDF and do regression in loglog scale, and increment the resulted *alpha* by 1. If constant width bins are used as in (a) we get $\alpha = 2.3533$. If logarithmic binning is used then $\alpha = 2.4567$.

(d) Using the MLE estimate we get $\alpha = 1 + n \left[\sum_{i=1}^n \ln \frac{x_i}{x_{min}} \right] = 2.4983$. ■

Problem 2.4 Consider the following evolving model for generating an undirected graph. Initially there are only three nodes connected into a triangle. At every time step, an edge of the current network is selected uniformly at random, and a new node is added to the network that links to both the endpoints of the edge. Prove that p_k , the fraction of nodes with degree k , follows a power law with exponent 3. Provide an intuitive explanation as to why this model is the same as the preferential attachment model.

Solution: Let $d_i(t)$ denote the degree of node i at time t . Node i only gets a new edge at time $t+1$ if one of his edges is picked. Hence, the expected value of $d_i(t+1)$ will be:

$$E[d_i(t+1)] = d_i(t) \cdot \left(1 + \frac{1}{3+2t}\right)$$

We can then approximate

$$\frac{\partial d_i(t)}{\partial t} \simeq \frac{d_i(t)}{3+2t}.$$

Solving the differential equation with the initial condition that $d_i(i) = 2$ we obtain

$$d_i(t) = 2 \left(\frac{3+2t}{3+2i} \right)^{\frac{1}{2}}.$$

Just as we did in class, we can now find which nodes have degree higher than k at time t :

$$i \leq \frac{2}{k^2}(3+2t) - \frac{3}{2}.$$

At time t there are $3+t$ nodes in the network, so the desired fraction is $p_k = \frac{2}{(3+t)k^2} \left((3+2t) - \frac{3}{2} \right)$. This expression can be considered the cdf (cumulative distribution function) of the degrees at time t . By derivating respect to k and making t tend to infinity, we get the asymptotic probability distribution:

$$p_k \simeq \frac{8}{k^3}$$

This model is the same as the preferential attachment because in both cases nodes the probability that a node gets a new edge is proportional to its current degree. ■

Problem 2.5 In this exercise we will study the distribution of words in the English language. The data consists of a list of all the words in a dictionary and a text version of “A tale of Two Cities” by Charles Dickens (found at project Gutenberg). In the later, we have removed punctuation, apostrophes, etc... keeping only the 26 characters in the alphabet and the space.

(a) Write a program that reads the list of words provided and plot a graph showing the number of words that there exist of lengths between 3 and 8 (you can discard all other words). How fast does such number increase?

(b) Using the novel “A Tale of Two Cities” as a representative sample, we now plot how frequently each words is used in the English language. Sort the words in the novel along the x axis from the most frequent to the least, and plot their number of appearances (many words in the dictionary will not be in the novel. You should not take those into account). Does it follow a power law? If so, find an approximation for the exponent.

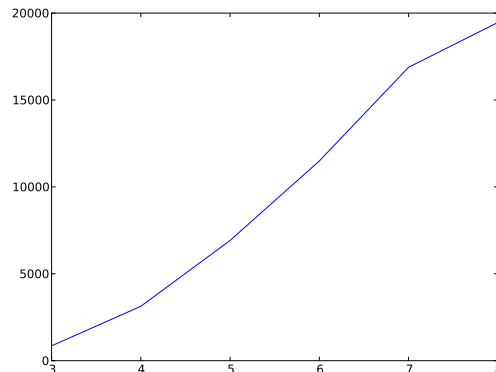
If you looked further into the previous plot, you would see that the most frequent words are usually shorter. We now develop models to explain why, if long words are more numerous in the dictionary, authors use short ones more often.

(c) Assume that a monkey typed one billion (10^9) random characters on a keyboard (26 letters + space bar), and call “word” any sequence of letters between two spaces. Find $f(n)$, the expected number of times that a GIVEN sequence of length n would appear in the monkey’s text (with spaces at both sides). Does $f(n)$ follow a power law? If so, find an approximation for the exponent.

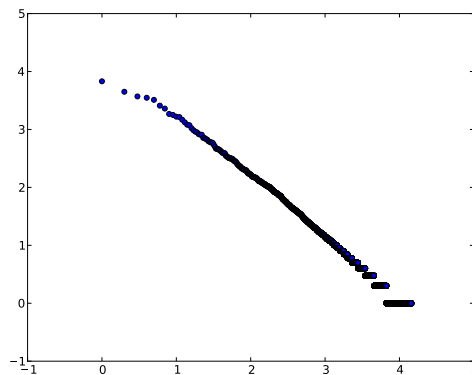
(d) In average, how many times would the 100-th most frequent word appear in the monkey’s text? What about the 1000-th? (Hint: how long would those words be? Either simulate it or find an analytic expression) Is this a good model for the results in (b)?

(e) We will try to further improve the model by assigning different probabilities to different characters. Find the probability of each character (including space) in “A Tale of Two Cities” and generate ten thousand words according to that distribution. Repeat the plot in part (b) for this new text. Is the model better?

Solution: (a) The number of words of a given length increases linearly between 3 and 8.



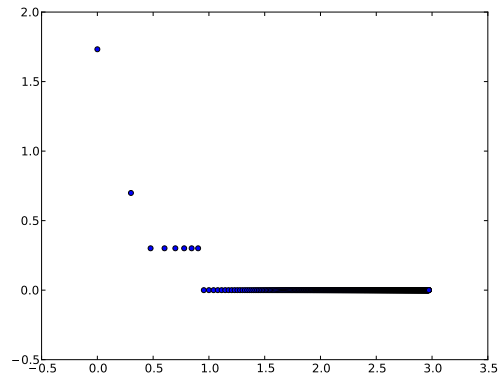
(b) Yes, it follows a power law, approximately with exponent -1.



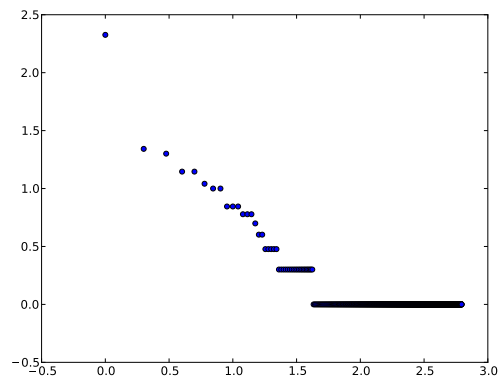
(c) Using the union bound, we get $f(n) = 10^9 \frac{26^n}{27^{n+2}}$. Rigorously speaking, it would be slightly smaller, since this is just an upper bound. It does not decrease according to a power law, but exponentially, as it becomes clear from the previous expression.

(d) In average, any two letter word will be more frequent than any three letter one, while two words with the same number of characters have the same chances of appearing. Therefore, the first 26 most frequent words will be 1-character ones. Then we will have the 26^2 two letter ones, which will roughly appear $f(2)$ times. Finally, the 1000^{th} most frequent word will have three characters, and appear with a frequency of $f(3)$.

It is not a good model for our data. It is too step-like. Although it is true that the two exponentials cancel each other (increasing number of words and decreasing frequency) giving a power law, it does not capture the progressive descent that we observed in (b).



(e) The model does improve. But there is still a large number of words that appear just once. By increasing the length of the randomly generated text we could improve the precision at the tail.



■